A Unified Grand Tour of Theoretical Physics

IAN D LAWRIE

IoP | SECOND EDITION

A Unified Grand Tour of Theoretical Physics

Second Edition

This page intentionally left blank

A Unified Grand Tour of Theoretical Physics

Second Edition

Ian D Lawrie

Reader in Theoretical Physics
The University of Leeds

Institute of Physics Publishing
Bristol and Philadelphia

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN 0 7503 0604 1

Library of Congress Cataloging-in-Publication Data are available

First Edition published 1990
First Edition reprinted 1994, 1998

Commissioning Editor: James Revill
Production Editor: Simon Laurenson
Production Control: Sarah Plenty
Cover Design: Frédérique Swist
Marketing Executive: Laura Serratrice

Published by Institute of Physics Publishing, wholly owned by The Institute of Physics, London

Institute of Physics Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

US Office: Institute of Physics Publishing, The Public Ledger Building, Suite 1035, 150 South Independence Mall West, Philadelphia, PA 19106, USA

Typeset in LaTeX 2ε by Text 2 Text, Torquay, Devon
Printed in the UK by MPG Books Ltd, Bodmin, Cornwall

Contents

This page intentionally left blank

Preface to the Second Edition

In preparing this revised edition of the Tour, I have corrected several errors and misprints for which I would like to take this opportunity of apologizing to readers of the first edition.

By now, supersymmetry and string theory have become so prominent in the theoretical physics literature (despite the more or less total absence of any experimental evidence of their relevance to the real world!) as to be obligatory in a book with this title. Accordingly, I have added introductory accounts of these topics in §12.7 and chapter 15. A comprehensive treatment of either topic (were I competent to write it) would require a book in itself, but I hope that the short accounts I have given will serve to make the extensive technical literature a little more accessible. I confess that I am no expert on string theory; Chris Hull and Jim Gates have given me advice which is perhaps enough to ensure that what I say is not grossly misleading, and I thank them for it.

Other new material in this edition includes a section on the applications of differential geometry to Newtonian mechanics and classical electromagnetism (§3.7) and a chapter on magnetic monopoles and other topological defects (chapter 13). I have also expanded my discussions of quantum fields in curved spacetimes (§7.7), grand unified theories (§12.6) and inflationary cosmology (§14.8) and attempted to improve and update my presentation of various other matters in minor ways.

I would like to thank IoP Publishing for giving me the opportunity of revising and extending the Tour. I am grateful to Jim Revill for his continual friendship and encouragement, and to Simon Laurenson for his unfailing patience and courtesy in dealing with the technicalities of bringing the final product into being.

Ian D Lawrie
October 2001

This page intentionally left blank

Preface to the First Edition

A few years ago, I decided to undertake some research having to do with the early history of the universe. It soon became apparent that I should have to improve my understanding of several aspects of theoretical physics, and it was from the ensuing process of self-education that the idea of writing this book emerged. I was particularly struck by two things. The first was the existence of many interrelationships, both physical and mathematical, between branches of physics that are traditionally treated as autonomous. The second was the lack of any textbook which had the scope to bring out these interrelationships adequately, or which would teach me at least the rudiments of what I needed to know in a relatively short time. It is that gap in the literature which I hope this book will go some way towards filling.

In trying to cover a wide range of topics, I have naturally been unable to give each the more extensive treatment it would receive in a more specialized work. I have tried to bear in mind the needs of three main categories of reader to whom I hope the book will be of use. As an undergraduate, I recall feeling annoying periods of frustration on encountering references to esoteric matters such as field theory and general relativity which were obviously important but said to be 'beyond the scope' of the lectures or recommended textbooks. Things have moved on a little since then, but it is still largely true that undergraduate courses devoted, for example, to gravitation and cosmology or elementary particle physics are required to give a broad view of the phenomenological aspects of their subjects, which leaves little room for exploring deeper aspects of their theoretical foundations. Final-year undergraduates who feel such a deprivation should find some enlightenment in these pages. Courses on 'theoretical physics' are also offered to undergraduates in physics and mathematics, perhaps as an optional alternative to some stint of laboratory work. The purpose of such a course is to illustrate the ways theoretical physicists have of thinking about the world, rather than to explore any of the subfields of physics exhaustively. I hope that this book will be found suitable as a basis for such courses, and have tried to arrange the material so that lecturers may select topics from it according to their own tastes.

Postgraduate students will no doubt find, as I have done, the need to acquire some familiarity with a wide range of material which is treated adequately only in rather forbidding technical treatises. They, I hope, will find here a palatable introduction to much of what they need and, indeed, a sufficient coverage of those topics which are peripheral to their chosen speciality.

Third, I have tried to provide for professional scientists and engineers who are not theoretical physicists. They, I conceive, may find themselves unsatisfied by semi-popular accounts of advances in the subject but without time for a full-scale assault on the technical literature. For them, this book may perhaps constitute a useful half-way house.

Responsibility for what appears herein is, of course, my own, but I should like to acknowledge the assistance I have received along the way. Much of what I understand of statistical mechanics was imparted some time ago by Michael Fisher. Others who have benefitted from his wisdom may recognize his influence in what I have to say, but he naturally bears no responsibility for anything I failed to understand properly. During 1986–7, I spent a sabbatical year at the University of British Columbia, where I had my first opportunity to teach a substantial graduate course on quantum field theory. The discipline of preparing the lectures and the perceptive response of the students who took the course did much to sharpen the somewhat less advanced presentation offered here. Euan Squires was instrumental in securing a contract for the book to be written. I have greatly appreciated his enthusiastic support during the writing and his comments on the first draft of the manuscript. I am also grateful to Gary Gibbons, who read the chapters on relativity and gravitation and saved me from a number of *faux pas*. Professor Jim Gates reviewed the entire manuscript, and I have greatly appreciated his many detailed comments and suggestions. It is a pleasure to thank Jim Revill, Neil Robertson and Jane Bartholomew at Adam Hilger for their assistance and encouragement during the various stages of production. The greatest thanks, perhaps, are due to my wife Ingrid who encouraged me through the whole venture and patiently allowed herself to be supplanted by textbooks and word processor through more evenings and weekends than either of us cares to remember.

Ian D Lawrie
December 1989

Glossary of Mathematical Symbols

∂_μ	partial derivative ($= \partial/\partial x^\mu$)	25
∇_μ	covariant derivative	32–3
$\Box$	d'Alembertian operator	141
$A_{,\mu}$	partial derivative	33
$A_{;\mu}$	covariant derivative	33
$A\overleftarrow{\partial}_\mu$	left-acting derivative	153
$A\overleftrightarrow{\partial}_\mu B$	antisymmetric derivative ($= A\partial_\mu B - (\partial_\mu A)B$)	142
$\mid\,\rangle$ ($\langle\,\mid$)	ket (bra) vector	112
A^{T}	transpose of a matrix A	
$\hat{A}$	operator in the Hilbert space of state vectors (in later chapters, the circumflex is omitted)	114
$\hat{A}^\dagger$	adjoint (or Hermitian conjugate) operator	115
$^*\boldsymbol{T}$	dual tensor	70
$\not{a}$	contraction with Dirac matrices ($= \gamma^\mu a_\mu$)	152
$\{A, B\}_{\mathrm{P}}$	Poisson bracket	54
$[\hat{A}, \hat{B}]$	commutator of two operators or matrices ($= AB - BA$)	35, 115
$\{A, B\}$	anticommutator of two matrices or operators ($= AB + BA$)	147
$S \otimes T$	tensor product	66
$\omega \wedge \sigma$	wedge product	67
$a(t)$	Robertson–Walker scale factor	380
A_μ	electromagnetic 4-vector potential	62
α	fine structure constant	227
β	inverse temperature ($= 1/k_{\mathrm{B}}T$)	243
c	fundamental speed	8
$C, \mathcal{C}$	charge conjugation matrices	156
γ^μ	Dirac matrices	147
γ^5	chirality matrix	152
γ_{ab}	worldsheet metric of a relativistic string	432
$\Gamma^\mu_{\nu\sigma}$	affine connection coefficients	31, 39
d	exterior derivative	70
$\mathrm{d}x^a$	basis one-form	66
$\delta_{ij}, \delta^{ij}, \delta^i_j$	Kronecker delta symbol	518

$\delta(x-y)$	Dirac delta function	518
e	fundamental charge	184
$e^{\mu}_{a}(x)$	vierbein	170
$F_{\mu\nu}$	field strength tensor	62, 187
$g_{\mu\nu}(x)$	metric tensor field	14, 36
$g(x)$	determinant of the metric tensor field	88
G	Newton's gravitational constant	11
G_{F}	Fermi constant	297
G_{F}, S_{F}, D_{F}	Feynman propagators	204, 210, 217
h	Planck's constant (also $\hbar = h/2\pi$)	108, 117
H	Hamiltonian	52
$\mathcal{H}$	Liouville operator	54
k_{B}	Boltzmann's constant	243
κ	gravitational constant ($= 8\pi G/c^4$)	88–90
L	Lagrangian	47
$\mathcal{L}$	Lagrangian density	63
L_n, $\widetilde{L}_n$	Virasoro generators	442–3
Λ	cosmological constant	88
$\Lambda^{\mu'}_{\ \mu}$	coordinate transformation matrix	26
$\mathbb{Q}$	configuration space	73
$\mathbb{R}$	the real line	16
$R^{\mu}_{\ \nu\sigma\tau}$	Riemann curvature tensor	35
$R_{\mu\nu}$	Ricci tensor	36
R	Ricci curvature scalar	39
ρ	phase-space probability density	54, 237
$\hat{\rho}$	density operator	251
S	action	47
σ^i	Pauli matrices	147
$T^{\mu\nu}$	stress tensor	60, 89
T_{ab}, T, $\widetilde{T}$	energy–momentum tensor of a relativistic string	433, 438
$T_P\mathbb{Q}$, $T^*_P\mathbb{Q}$	tangent and cotangent spaces	74–5
$T\mathbb{Q}$, $T^*\mathbb{Q}$	tangent and cotangent bundles	74–5
$T[\ldots]$	time ordered product	203
τ	proper time	10
$\eta_{\mu\nu}$	metric tensor of Minkowski spacetime	14
$\Psi(\boldsymbol{x}, t)$	wavefunction	109
$X^{\mu}(\tau, \sigma)$	spacetime coordinate of the point (τ, σ) on the worldsheet of a relativistic string	431
z	fugacity	244
Z_{can}, Z_{gr}	canonical and grand canonical partition functions	242, 244
$\omega^{a}_{\ b\nu}$	spin connection	171
Ω	symplectic 2-form	76
$\Omega(t)$	cosmological density ratio	387

Chapter 1

Introduction: The Ways of Nature

In the eighteenth century, it became fashionable for wealthy young Englishmen to undertake the Grand Tour, an excursion which may have lasted several years, their principal destinations being Paris and the great cultural centres of Italy—Rome, Venice, Florence and Naples. For many, no doubt, the joys of traveling and occasional revelry were a sufficient inducement. For others, the opportunity to observe at first hand the social, literary and artistic achievements of other nations represented the completion of their liberal education. For a few, perhaps, it was the starting point of an independent intellectual career. It is in somewhat the same spirit that I wish to offer readers of this book a guided grand tour of theoretical physics. The members of my party need be neither wealthy (my publisher permitting), young, English nor male. I am, however, going to assume that they have a sound knowledge of basic physics, such as a student in his or her final year of undergraduate study ought to possess.

Our itinerary cannot, of course, include everything that is important in theoretical physics. Our principal destinations are those central ideas which form the foundations of our understanding of how the world works—our knowledge, as it now stands, of the ways of nature. In outline, the topics I plan to explore are: the theories of relativity, which concern themselves with the geometrical structure of space and time and from which emerge an account of gravitational phenomena; quantum mechanics and quantum field theory, which describe the constitution of matter at the most microscopic level that is currently accessible to experiments; and statistical mechanics, which, up to a point, allows us to deduce from this microscopic constitution the properties of the macroscopic systems of which the universe is principally composed. The universe itself, and especially its early history, form the subject of the penultimate chapter, where many of the ideas we shall have explored must be brought into play. In the final chapter, I give an introduction to the more speculative theory of quantized relativistic strings (and, as it turns out, of other objects too) which, in the eyes of its advocates at least, promises to provide the most comprehensive account it has so far been possible to devise of the ways of nature at the most fundamental level.

For some readers, the desire to gain a little insight into our contemporary understanding of the ways of nature will, I hope, be a sufficient inducement to read this book. For others, such as those nearing the end of their undergraduate studies, I hope to provide the opportunity of rounding off that stage of their education by delving a little more deeply into the ways of nature than the core of an undergraduate curriculum normally does. For a few, such as those embarking upon postgraduate research in fundamental theoretical physics, I hope to provide a readily digestible introduction to many of the ideas that they will need to master.

Before setting out, I should say a few words about the point of view from which the book is written. By and large, I have written only about what I know and what I believe I understand. This, and the limited number of pages at my disposal, have led to the omission of many topics that other writers might consider essential to a theoretical understanding of physics, but that cannot be helped. The topics I have included are those that I believe to be fundamental, in the sense that I have tried to convey by speaking of the 'ways of nature'. The philosopher Karl Popper would have us believe that scientific theories exist only to be refuted by experimental evidence. If practising scientists really thought in that way, then I doubt that they would consider their expenditure of intellectual effort worthwhile. A good scientific theory is seldom refuted by new experimental evidence for which it cannot account. Much more often, it comes to be extended, generalized or reinterpreted as a constituent part of some more comprehensive theory. Every time this happens, we improve our understanding of what the world is really like: we gain a clearer picture of the ways of nature.

The way in which such transformations in our understanding come about is not necessarily apparent at the point where a detailed theoretical prediction is confronted with an experimental datum. Take, for example, the transformation of classical Newtonian mechanics into quantum mechanics. We have discovered, amongst other things, that electrons can be diffracted by crystals: a phenomenon for which quantum mechanics can account but classical mechanics cannot. Therefore, it is often said, classical mechanics must be wrong, or at least no more than an approximation to quantum mechanics with a restricted range of usefulness. It is indeed true that, under appropriate circumstances, the predictions of classical mechanics can be regarded as a good approximation to those of quantum mechanics, but that is the less interesting part of the truth. There is, as we shall see, a level of description (which is not especially esoteric) at which classical and quantum mechanics are virtually identical, apart from a change of interpretation, and it is the reinterpretation that is vital and profound. It is, I maintain, at such a level of description that an understanding of the ways of nature is to be sought, and it is that level of description that is emphasized in this book.

It would, of course, be absurd to lay claim to any understanding of the ways of nature if our theories could not be tested in detail against experimental observations. Unfortunately, the task of deriving from our fundamental theories precise predictions that can be subjected to stringent experimental tests is often a long and highly technical one. This task, like the devising of the experiments themselves, is essential and intellectually challenging but, for want of the necessary space, I shall not often describe in detail how it can be accomplished. I do not think that this requires any apology. The basic conceptual understanding I hope to provide can, on first acquaintance, be obscured by the technical details of specific applications. Readers will nevertheless want to know by what right the theories I present can claim to describe the ways of nature, and I shall indeed outline, at certain key points, the evidence on which this claim is based. Readers who wish to become professional physicists will, in the end, have to master at least those details that are relevant to their chosen speciality and will find them described in many excellent, specialized textbooks, some of which are mentioned in my bibliography.

Most good scientific theories have been born of the need to understand certain puzzling observations. If, in retrospect, our improved insight into the ways of nature shows us that those observations are no longer puzzling but entirely to be expected, then we feel satisfied that the desired understanding has been achieved. We feel this satisfaction most deeply when the theory we have constructed has a coherent, logical, aesthetically pleasing internal structure, and rests on a few basic assumptions which, though they may not be quite self-evident, have a convincing ring of truth. Almost, though never entirely, we come to feel that things could not really have been any other way. It may be presumptuous to suppose that the ways of nature must necessarily have such a psychological appeal for us. The fact is, though, that the most successful fundamental theories of physics are of this kind, and that, for me and many others, is what makes the enterprise worthwhile.

My desire to bring out this aspect of theoretical physics strongly influences the way this book is written. When discussing, in particular, relativity and quantum mechanics, the main part of my treatment begins by describing the theoretical concepts and mathematical structures that lie at the heart of these theories, and later develops some of their consequences in particular physical situations. The more traditional method of introducing these subjects is to set out at the beginning the experimental facts that stand in need of explanation and then to ask what new theoretical concepts are needed to accommodate them. I realize that, for many readers, the traditional approach is the more easily accessible one. For that reason, I have given in §§2.0 and 5.0 short summaries of the more traditional development of elementary aspects of the theory. To some extent, these should serve as previews of the more detailed accounts that follow and enable readers to preserve a sense of direction and purpose while the mathematical formalism is developed. Ideally, readers should already be acquainted with special relativity, the wave-mechanical version of quantum mechanics and their simpler applications. Readers who are thus equipped may prefer to skip these introductory sections or to regard them and the more elementary exercises as a short revision course.

In the main, my treatment of mathematical formalism is intended to be complete and explicit. Where I have omitted the algebraic details needed to derive an equation, readers should be able to supply them, and should usually not be satisfied until they have done so. In some cases, the exercises offer guidance. The exercises should, indeed, be regarded as an integral part of the tour; some of them introduce important ideas that are not dealt with fully in the main text. Occasionally, it is necessary for me merely to quote the result of a calculation that is too lengthy or technical to be reproduced in detail, and I shall indicate when this is so.

There is one other aspect of theoretical physics that I should like readers to be aware of. It has become apparent that there are many similarities, some of them physical and others mathematical, between areas of physics which, on the face of it, appear to be quite separate. In the course of this book, I emphasize two of these unifying themes particularly. One is that the geometrical ideas we need to describe the structure of space and time also lie at the root of the gauge theories of fundamental forces, described in chapters 8 and 12, of which the most familiar is electromagnetism. Indeed, once we realize the importance of these ideas, the existence of both gravitational and other forces is seen to be almost inevitable, even if we had not already been aware of them. The other is a basic mathematical similarity between quantum field theory and statistical mechanics which, as I discuss in chapter 10, can appear in several different guises. This is not altogether surprising, since both theories require us to average over uncertainties of one kind or another. The extent of the similarity is, however, quite striking, and becomes particularly apparent in the study of phase transitions, with which I deal in chapter 11. One of my chief ambitions in writing this book is to offer a unified account of theoretical physics in which these interconnections can properly be brought out.

While the connections between different topics will be appreciated only by those who read the book in its entirety, I have tried to arrange the material so that not all of it need be mastered in one go. Readers who are mainly interested in relativity and gravitation may read chapters 2, 3 and 4 and the first three sections of chapter 14 without serious loss of continuity, though the remainder of chapter 14 requires some knowledge of particle physics and statistical mechanics. Similarly, those whose main interest is in particles and field theory may read chapters 3, 5–9 and 12, together with the more speculative material of chapters 13 and 15, but should preferably look at §§2.0 and 11.4–11.7 for some background information. They should then be able to follow most of chapter 14. Chapters 3, 5, 10 and 11 can be read as a short course on statistical physics and the theory of phase transitions. Readers who follow one of these schemes may safely ignore occasional references to unfamiliar material, or may like to dip into relevant portions of the chapters they have omitted.

The purpose of this book is entirely pedagogical. I do not aim to describe the history of theoretical physics, nor to give anything approaching a comprehensive survey of the research literature. As far as possible, I have made at least passing mention of important ideas which bear on the topics I discuss but cannot be covered in detail, and the bibliography lists a number of good textbooks and review articles to which interested readers may turn for further information and references to the original literature. I have given some references to the literature where I think that readers will find an original paper particularly enlightening or where it provides a useful historical perspective, but I have by no means listed every paper in these categories. I have certainly not attempted to refer explicitly to the work of every scientist who has made important contributions to the subjects I discuss. To do so would require a book in itself.

It is time for our tour to begin.

Chapter 2

Geometry

Our tour of theoretical physics begins with geometry, and there are two reasons for this. One is that the framework of space and time provides, as it were, the stage upon which physical events are played out, and it will be helpful to gain a clear idea of what this stage looks like before introducing the cast. As a matter of fact, the geometry of space and time itself plays an active role in those physical processes that involve gravitation (and perhaps, according to some speculative theories, in other processes as well). Thus, our study of geometry will culminate, in chapter 4, in the account of gravity offered by Einstein's general theory of relativity. The other reason for beginning with geometry is that the mathematical notions we develop will reappear in later contexts.

To a large extent, the special and general theories of relativity are 'negative' theories. By this I mean that they consist more in relaxing incorrect, though plausible, assumptions that we are inclined to make about the nature of space and time than in introducing new ones. I propose to explain how this works in the following way. We shall start by introducing a prototype version of space and time, called a 'differentiable manifold', which possesses a bare minimum of geometrical properties—for example, the notion of length is not yet meaningful. (Actually, it may be necessary to abandon even these minimal properties if, for example, we want a geometry that is fully compatible with quantum theory and I shall touch briefly on this in chapter 15.) In order to arrive at a structure that more closely resembles space and time as we know them, we then have to endow the manifold with additional properties, known as an 'affine connection' and a 'metric'. Two points then emerge: first, the common-sense notions of Euclidean geometry correspond to very special choices for these affine and metric properties; second, other possible choices lead to geometrical states of affairs that have a natural interpretation in terms of gravitational effects. Stretching the point slightly, it may be said that, merely by *avoiding* unnecessary assumptions, we are able to see gravitation as something entirely to be expected, rather than as a phenomenon in need of explanation.

To me, this insight into the ways of nature is immensely satisfying, and it is in the hope of communicating this satisfaction to readers that I have chosen to approach the subject in this way. Unfortunately, the assumptions we are to avoid are, by and large, *simplifying* assumptions, so by avoiding them we let ourselves in for some degree of complication in the mathematical formalism. Therefore, to help readers preserve a sense of direction, I will, as promised in chapter 1, provide an introductory section outlining a more traditional approach to relativity and gravitation, in which we ask how our naïve geometrical ideas must be modified to embrace certain observed phenomena.

2.0 The Special and General Theories of Relativity

2.0.1 The special theory

The special theory of relativity is concerned in part with the relation between observations of some set of physical events in two inertial frames of reference that are in relative motion. By an inertial frame, we mean one in which Newton's first law of motion holds:

> Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed on it.
> (Newton 1686)

It is worth noting that this definition by itself is in danger of being a mere tautology, since a 'force' is in effect defined by Newton's second law in terms of the acceleration it produces:

> The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.
> (Newton 1686)

So, from these definitions alone, we have no way of deciding whether some observed acceleration of a body relative to a given frame should be attributed, on the one hand, to the action of a force or, on the other hand, to an acceleration of the frame of reference. Eddington has made this point by a facetious re-rendering of the first law:

> Every body tends to move in the track in which it actually does move, except insofar as it is compelled by material impacts to follow some other track than that in which it would otherwise move.
> (Eddington 1929)

The extra assumption we need, of course, is that forces can arise only from the influence of one body on another. An inertial frame is one relative to which any body sufficiently well isolated from all other matter for these influences to be negligible does not accelerate. In practice, needless to say, this isolation cannot be achieved. The successful application of Newtonian mechanics depends on our being able systematically to identify, and take proper account of, all those forces

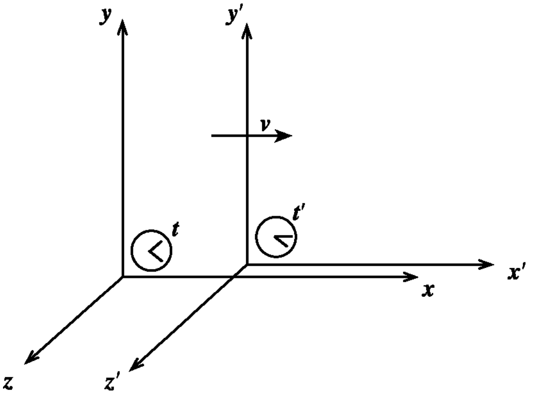


Figure 2.1. Two systems of Cartesian coordinates in relative motion.

that cannot be eliminated. To proceed, we must take it as established that, in principle, frames of reference can be constructed, relative to which any isolated body will, as a matter of fact, always refuse to accelerate. These frames we call inertial.

Obviously, any two inertial frames must either be relatively at rest or have a uniform relative velocity. Consider, then, two inertial frames, S and S' (standing for Systems of coordinates) with Cartesian axes so arranged that the x and x' axes lie in the same line, and suppose that S' moves in the positive x direction with speed v relative to S. Taking y' parallel to y and z' parallel to z, we have the arrangement shown in figure 2.1. We assume that the sets of apparatus used to measure distances and times in the two systems are identical and, for simplicity, that both clocks are adjusted to read zero at the moment the two origins coincide.

Suppose that an event at the coordinates (x, y, z, t) relative to S is observed at (x', y', z', t') relative to S'. According to the Galilean, or common-sense, view of space and time, these two sets of coordinates must be related by

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t. \tag{2.1}$$

Since the path of a moving particle is just a sequence of events, we easily find that its velocity relative to S, in vector notation $\boldsymbol{u} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$, is related to its velocity $\boldsymbol{u}' = \mathrm{d}\boldsymbol{x}'/\mathrm{d}t'$ relative to S' by $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$, with $\boldsymbol{v} = (v, 0, 0)$, and that its acceleration is the same in both frames, $\boldsymbol{a}' = \boldsymbol{a}$.

Despite its intuitive plausibility, the common-sense view turns out to be mistaken in several respects. The special theory of relativity hinges on the fact that the relation $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$ is not true. That is to say, this relation disagrees with experimental evidence, although discrepancies are detectable only when speeds are involved whose magnitudes are an appreciable fraction of a fundamental speed c, whose value is approximately $2.998 \times 10^8\,\mathrm{m\,s^{-1}}$. So far as is known, light travels through a vacuum at this speed, which is, of course, generally called the speed of light. Indeed, the speed of light is predicted by Maxwell's electromagnetic theory to be $(\epsilon_0\mu_0)^{-1/2}$ (in SI units, where ϵ_0 and μ_0 are called the permittivity and permeability of free space, respectively) but the theory does not single out any special frame relative to which this speed should be measured. For quite some time after the appearance of Maxwell's theory (published in its final form in 1864; see also Maxwell (1873)), it was thought that electromagnetic radiation consisted of vibrations of a medium, the 'luminiferous ether', and would travel at the speed c relative to the rest frame of the ether. However, a number of experiments cast doubt on this interpretation. The most celebrated, that of Michelson and Morley (1887), showed that the speed of the Earth relative to the ether must, at any time of year, be considerably smaller than that of its orbit round the Sun. Had the ether theory been correct, of course, the speed of the Earth relative to the ether should have changed by twice its orbital speed over a period of six months. The experiment seemed to imply, then, that light always travels at the same speed, c, relative to the apparatus used to observe it.

In his paper of 1905, Einstein makes the fundamental assumption (though he expresses things a little differently) that *light travels with exactly the same speed, c, relative to any inertial frame*. Since this is clearly incompatible with the Galilean transformation law given in (2.1), he takes the remarkable step of modifying this law to read

$$\begin{aligned} x' &= \frac{x - vt}{(1 - v^2/c^2)^{1/2}} & y' &= y \\ z' &= z & t' &= \frac{t - vx/c^2}{(1 - v^2/c^2)^{1/2}}. \end{aligned} \tag{2.2}$$

These equations are known as the *Lorentz transformation*, because a set of equations having essentially this form had been written down by H A Lorentz (1904) in the course of his attempt to explain the results of Michelson and Morley. However, Lorentz believed that his equations described a mechanical effect of the ether upon bodies moving through it, which he attributed to a modification of intermolecular forces. He does not appear to have interpreted them as Einstein did, namely as a general law relating coordinate systems in relative motion. The assumptions that lead to this transformation law are set out in exercise 2.1, where readers are invited to complete its derivation. Here, let us note that (2.2) does indeed embody the assumption that light travels with speed c relative to any inertial frame. For example, if a pulse of light is emitted from the common origin of S and S' at $t = t' = 0$, then the equation of the resulting spherical wavefront at time t relative to S is $x^2 + y^2 + z^2 = c^2t^2$. Using the transformation (2.2), we easily find that its equation at time t' relative to S' is $x'^2 + y'^2 + z'^2 = c^2t'^2$.

Many of the elementary consequences of special relativity follow directly from the Lorentz transformation, and we shall meet some of them in later chapters. What particularly concerns us at present—and what makes Einstein's interpretation of the transformation equations so remarkable—is the change that these equations require us to make in our view of space and time. On the face of it, equations (2.1) or (2.2) simply tell us how to relate observations made in two different frames of reference. At a deeper level, however, they contain information about the structure of space and time that is independent of any frame of reference. Consider two events with spacetime coordinates $(\boldsymbol{x}_1, t_1)$ and $(\boldsymbol{x}_2, t_2)$ relative to S. According to the Galilean transformation, the time interval $t_2 - t_1$ between them relative to S is equal to the interval $t'_2 - t'_1$ relative to S'. In particular, it may happen that these two events are simultaneous, so that $t_2 - t_1 = 0$, and this statement would be equally valid from the point of view of either frame of reference. For two simultaneous events, the spatial distances between them, $|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ and $|\boldsymbol{x}'_1 - \boldsymbol{x}'_2|$ are also equal. Thus, the time interval between two events and the spatial distance between two simultaneous events have the same value in *every* inertial frame, and hence have real physical meanings that are independent of any system of coordinates. According to the Lorentz transformation (2.2), however, both the time interval and the distance have different values relative to different inertial frames. Since these frames are arbitrarily chosen by us, neither the time interval nor the distance has any definite, independent meaning. The one quantity that does have a definite, frame-independent meaning is the *proper time interval* $\Delta\tau$, defined by

$$c^2\Delta\tau^2 = c^2\Delta t^2 - \Delta x^2 \tag{2.3}$$

where $\Delta t = t_2 - t_1$ and $\Delta x = |\boldsymbol{x}_2 - \boldsymbol{x}_1|$. By using (2.2), it is easy to verify that $c^2\Delta t'^2 - \Delta x'^2$ is also equal to $c^2\Delta\tau^2$.

We see, therefore, that the Galilean transformation can be correct only in a *Galilean spacetime*; that is, a spacetime in which both time intervals and spatial distances have well-defined meanings. For the Lorentz transformation to be correct, the structure of space and time must be such that only proper-time intervals are well defined. There are, as we shall see, many such structures. The one in which the Lorentz transformation is valid is called *Minkowski spacetime* after Hermann Minkowski who first clearly described its geometrical properties (Minkowski, 1908). These properties are summarized by the definition (2.3) of proper time intervals. In this definition, the constant c does not refer to the speed of anything. Although it has the dimensions of velocity, its role is really no more than that of a conversion factor between units of length and time. Thus, although the special theory of relativity arose from attempts to understand the propagation of light, it has nothing to do with electromagnetic radiation as such. Indeed, it is not in essence about relativity either! Its essential feature is the structure of space and time expressed by (2.3), and the law for transforming between frames in relative motion serves only as a clue to what this structure is. With this in mind, Minkowski (1908) says of the name ‘relativity’ that it ‘... seems to me very feeble’.

The geometrical structure of space and time restricts the laws of motion that may govern the dynamical behaviour of objects that live there. This is true, at least, if one accepts the *principle of relativity*, expressed by Einstein as follows:

> The laws by which the states of physical systems undergo change are not affected, whether these changes of state be referred to the one or the other of two systems of coordinates in uniform translatory motion.
> (Einstein 1905)

Any inertial frame, that is to say, should be as good as any other as far as the laws of physics are concerned. Mathematically, this means that the equations expressing these laws should be *covariant*—they should have the same form in any inertial frame. Consider, for example, two objects, with masses m_1 and m_2, situated at x_1 and x_2 on the x axis of S. According to Newtonian mechanics and the Newtonian theory of gravity, the equation of motion for particle 1 is

$$m_1 \frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} = (Gm_1m_2)\frac{x_2 - x_1}{|x_2 - x_1|^3} \tag{2.4}$$

where $G \simeq 6.67 \times 10^{-11}\,\mathrm{N\,m^2\,kg^{-2}}$ is Newton's gravitational constant. If spacetime is Galilean and the transformation law (2.1) is valid, then $\mathrm{d}^2x'/\mathrm{d}t'^2 = \mathrm{d}^2x/\mathrm{d}t^2$ and $(x'_2 - x'_1) = (x_2 - x_1)$, so in S' the equation has exactly the same form and Einstein's principle is satisfied. In Minkowski spacetime, we must use the Lorentz transformation. The acceleration relative to S is not equal to the acceleration relative to S' (see exercise 2.2), but worse is to come! On the right-hand side, x_1 and x_2 refer to two events, namely the objects reaching these two positions, which occur simultaneously as viewed from S. As viewed from S', however, these two events are separated by a time interval $(t'_2 - t'_1) = (x'_1 - x'_2)v/c^2$, as readers may easily verify from (2.2). In Minkowski spacetime, therefore, (2.4) does not respect the principle of relativity. It is unsatisfactory as a law of motion because it implies that there is a preferred inertial frame, namely S, relative to which the force depends only on the instantaneous separation of the two objects; relative to any other frame, it depends on the distance between their positions at different times, and also on the velocity of the frame of reference relative to the preferred one. Actually, we do not know *a priori* that there is no such preferred frame. In the end, we trust the principle of relativity because the theories that stem from it explain a number of observed phenomena for which Newtonian mechanics cannot account.

We might imagine that electrical forces would present a similar problem, since we obtain Coulomb's law for particles with charges q_1 and q_2 merely by replacing the constant in parentheses in (2.4) with $-q_1q_2/4\pi\epsilon_0$. In fact, Maxwell's theory is not covariant under Galilean transformations, but can be made covariant under Lorentz transformations with only minor modifications. We shall deal with electromagnetism in some detail later on, and I do not want to enter into the technicalities at this point. We may note, however, the features that favour Lorentz covariance. In Maxwell's theory, the forces between charged particles are transmitted by electric and magnetic fields. We know that the fields due to a charged particle do indeed appear different in different inertial frames: in a frame in which the particle is at rest, we see only an electric field, while in a frame in which the particle is moving, we also see a magnetic field. Moreover, disturbances in these fields are transmitted at the speed of light. The problem of simultaneity is avoided because a second particle responds not directly to the first one, but rather to the electromagnetic field at its own position. The expression analogous to the right-hand side of (2.4) for the Coulomb force is valid only when there is a frame of reference in which particle 2 can be considered fixed, and then only as an approximation.

2.0.2 The general theory

The experimental fact that eventually led to the special theory was, as we have seen, the constancy of the speed of light. The general theory, and the account that it provides of gravitation, also spring from a crucial fact of observation, namely the equality of inertial and gravitational masses. In (2.4), the mass m_1 appears in two different guises. On the left-hand side, m_1 denotes the *inertial mass*, which governs the response of the body to a given force. On the right-hand side, it denotes the *gravitational mass*, which determines the strength of the gravitational force. The gravitational mass is analogous to the electric charge in Coulomb's law and, since the electrical charge on a body is not necessarily proportional to its mass, there is no obvious reason why the gravitational 'charge' should be determined by the mass either. The equality of gravitational and inertial masses is, of course, responsible for the fact that the acceleration of a body in the Earth's gravitational field is independent of its mass, and this has been familiar since the time of Galileo and Newton. It was checked in 1889 to an accuracy of about one part in 10^9 by Eötvös, whose method has been further refined more recently by R H Dicke and his collaborators.

It seemed to Einstein that this precise equality demanded some explanation, and he was struck by the fact that *inertial forces* such as centrifugal and Coriolis forces are proportional to the inertial mass of the body on which they act. These inertial forces are often regarded as 'fictitious', in the sense that they arise from the use of accelerating (and therefore non-inertial) frames of reference. Consider, for example, a spaceship far from any gravitating bodies such as stars or planets. When its motors are turned off, a frame of reference S fixed in the ship is inertial provided, as we assume, that it is not spinning relative to distant stars. Relative to this frame, the equation of motion of an object on which no forces act is $m\mathrm{d}^2\boldsymbol{x}/\mathrm{d}t^2 = 0$. Suppose the motors are started at time $t = 0$, giving the ship a constant acceleration a in the x direction. S is now not an inertial frame. If S' is the inertial frame that coincided with S for $t < 0$, then the equation of the object is still $m\mathrm{d}^2\boldsymbol{x}'/\mathrm{d}t'^2$=0, at least until the object collides with the cabin walls. Using Galilean relativity for simplicity, we have $x' = x + \frac{1}{2}at^2$ and $t' = t$, so relative to S the equation of motion is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -ma. \tag{2.5}$$

The force on the right-hand side arises trivially from the coordinate transformation and is definitely proportional to the *inertial* mass.

Einstein's idea is that gravitational forces are of essentially the same kind as that appearing in (2.5), which means that the inertial and gravitational masses are necessarily identical. Suppose that the object in question is in fact a physicist, whose ship-board laboratory is completely soundproof and windowless. His sensation of weight, as expressed by (2.5), is equally consistent with the ship's being accelerated by its motors or with its having landed on a planet at whose surface the acceleration due to gravity is a. Conversely, when he was apparently weightless, he would be unable to tell whether his ship was actually in deep space or freely falling towards a nearby planet. This illustrates Einstein's *principle of equivalence*, according to which the effects of a gravitational field can locally be eliminated by using a freely-falling frame of reference. This frame is inertial and, relative to it, the laws of physics take the same form that they would have relative to any inertial frame in a region far removed from any gravitating bodies.

The word 'locally' indicates that the freely-falling inertial frame can usually extend only over a small region. Let us suppose that our spaceship is indeed falling freely towards a nearby planet. (Readers may rest assured that the pilot, unlike the physicist, is aware of this and will eventually act to avert the impending disaster.) If he has sufficiently accurate apparatus, the physicist can detect the presence of the planet in the following way. Knowing the standard landing procedure, he allows two small objects to float freely on either side of his laboratory, so that the line joining them is perpendicular to the direction in which he knows that the planet, if any, will lie. Each of these objects falls towards the centre of the planet, and therefore their paths slowly converge. As observed in the freely-falling laboratory, they do not accelerate in the direction of the planet, but they do accelerate towards each other, even though their mutual gravitational attraction is negligible. (The tendency of the cabin walls to converge in the same manner is, of course, counteracted by interatomic forces within them.) Strictly, then, the effects of gravity are eliminated in the freely-falling laboratory only to the extent that two straight lines passing through it, which meet at the centre of the planet, can be considered parallel. If the laboratory is small compared with its distance from the centre of the planet, then this will be true to a very good approximation, but the equivalence principle applies exactly only to an infinitesimal region.

The principle of equivalence as stated above is not as innocuous as it might appear. We illustrated it by considering the behaviour of freely-falling objects, and found that it followed in a more or less trivial manner from the equality of gravitational and inertial masses. A version restricted to such situations is sometimes called the *weak* principle of equivalence. The *strong* principle, applying to all the laws of physics, has much more profound implications. It led Einstein to the view that gravity is not a force of the usual kind. Rather, the effect of a massive body is to modify the geometry of space and time. Particles that are not acted on by any ordinary force do not accelerate; they merely appear to be accelerated by gravity if we make the false assumption that the geometry is that of Galilean or Minkowski spacetime and interpret our observations accordingly.

Consider again the expression for proper time intervals given in (2.3). It is valid when (x, y, z, t) refer to Cartesian coordinates in an inertial frame of reference. In the neighbourhood of a gravitating body, a freely-falling inertial frame can be defined only in a small region, so we write it as

$$c^2(\mathrm{d}\tau)^2 = c^2(\mathrm{d}t)^2 - (\mathrm{d}\boldsymbol{x})^2 \tag{2.6}$$

where $\mathrm{d}t$ and $\mathrm{d}\boldsymbol{x}$ are infinitesimal coordinate differences. Now let us make a transformation to an arbitrary system of coordinates (x^0, x^1, x^2, x^3), each new coordinate being expressible as some function of x, y, z and t. Using the chain rule, we find that (2.6) becomes

$$c^2(\mathrm{d}\tau)^2 = \sum_{\mu,\nu=0}^{3} g_{\mu\nu}(x)\mathrm{d}x^\mu \mathrm{d}x^\nu \tag{2.7}$$

where the functions $g_{\mu\nu}(x)$ are given in terms of the transformation functions. They are components of what is called the *metric tensor*. In the usual version of general relativity, it is the metric tensor that embodies all the geometrical structure of space and time. Suppose we are given a set of functions $g_{\mu\nu}(x)$ which describe this structure in terms of some system of coordinates $\{x^\mu\}$. According to the principle of equivalence, it is possible at any point (say X, with coordinates X^μ) to construct a freely falling inertial frame, valid in a small neighbourhood surrounding X, relative to which there are no gravitational effects and all other processes occur as in special relativity. This means that it is possible to find a set of coordinates (ct, x, y, z) such that the proper time interval (2.7) reverts to the form of (2.6). Using a matrix representation of the metric tensor, we can write

$$g_{\mu\nu}(X) = \eta_{\mu\nu} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{2.8}$$

where $\eta_{\mu\nu}$ is the special metric tensor corresponding to (2.6).

If the geometry is that of Minkowski spacetime, then it will be possible to choose (ct, x, y, z) in such a way that $g_{\mu\nu} = \eta_{\mu\nu}$ everywhere. Otherwise, the best we can usually do is to make $g_{\mu\nu} = \eta_{\mu\nu}$ at a single point (though that point can be anywhere) or at every point along a curve, such as the path followed by an observer. Even when we do not have a Minkowski spacetime, it may be possible to set up an approximately inertial and approximately Cartesian coordinate system such that $g_{\mu\nu}$ differs only a little from $\eta_{\mu\nu}$ throughout a large region. In such a case, we can do much of our physics successfully by assuming that spacetime is exactly Minkowskian. If we do so, then, according to general relativity, we shall interpret the slight deviations from the true Minkowski metric as gravitational forces.

This concludes our introductory survey of the theories of relativity. We have concentrated on the ways in which our common-sense ideas of spacetime geometry must be modified in order to accommodate two key experimental observations: the constancy of the speed of light and the equality of gravitational and inertial masses. It is clear that the modified geometry leads to modifications in the laws that govern the behaviour of physical systems, but we have not discussed these laws in concrete terms. That we shall be better equipped to do after we have developed some mathematical tools in the remainder of this chapter. At that stage, we shall be able to see much more explicitly how gravity arises from geometry.

2.1 Spacetime as a Differentiable Manifold

Our aim is to construct a mathematical model of space and time that involves as few assumptions as possible, and to be explicitly aware of the assumptions we do make. In particular, we have seen that the theories of relativity call into question the meanings we attach to distances and time intervals, and we need to be clear about these. The mathematical structure that has proved to be a suitable starting point, at least for a non-quantum-mechanical model of space and time, is called a *differentiable manifold*. It is a collection of *points*, each of which will eventually correspond to a unique position in space and time, and the whole collection comprises the entire history of our model universe. It has two key features that represent familiar facts about our experience of space and time. The first is that any point can be uniquely specified by a set of four real numbers, so spacetime is four-dimensional. For the moment, the exact number of dimensions is not important. Later on, indeed, we shall encounter some recent theories which suggest that there may be more than four, the extra ones being invisible to us. Even in more conventional theories, we shall find that it is helpful to consider other numbers of dimensions as a purely mathematical device. The second feature is a kind of ‘smoothness’, meaning roughly that, given any two distinct points, there are more points in between them. This feature allows us to describe physical quantities such as particle trajectories or electromagnetic fields in terms of differentiable functions and hence to do theoretical physics of the usual kind. We do not know for certain that space and time are quite as smooth as this, but at least there is no evidence for any granularity down to the shortest distances we are able to probe experimentally.

Our first task is to express these properties in a more precise mathematical form. It is of fundamental importance that this can be done without recourse to any notion of length. The properties we require are *topological* ones, and we begin by introducing some elementary ideas of topology. Roughly speaking, we want to be able to say that some pairs of points are ‘closer together’ than others, without having any quantitative measure of distance. As an illustration, consider a sheet of rubber, marked off into different regions as in figure 2.2. For the purposes of this illustration, we shall say that there is no definite distance between two points

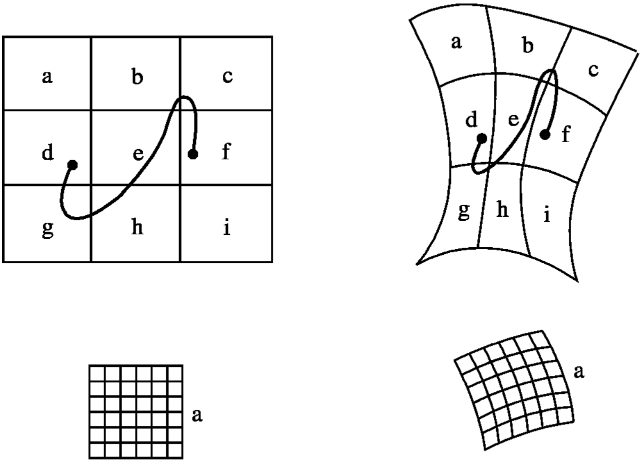
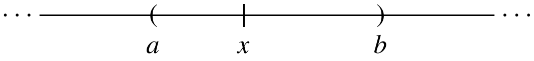


Figure 2.2. A deformable sheet of rubber, divided into several regions. Although there is no definite distance between the points indicated by ● , there are always other points between them, because any curve joining them must pass through at least one of the regions b, e and h.

on the sheet, because it can be deformed at will. No matter how it is deformed, however, any given region is still surrounded by the same neighbouring regions. Given a point in d and another in f, we can never draw a line between them that does not pass through at least one of regions b, e and h. The same holds, moreover, of more finely subdivided regions, as shown for subdivisions of a, each of which could be further subdivided, and so on. In this sense, points on the sheet are smoothly connected together. The smoothness would be lost if the rubber were vaporized, the individual molecules being considered as the collection of points. Mathematically, the kind of smoothness we want is a property of the real line (that is, the set of all real numbers, denoted by $\mathbb{R}$). So, as part of the definition of the manifold, we demand that it should be possible to set up correspondences (called 'maps') between points of the manifold and sets of real numbers. We shall next look at the topological properties of real numbers, and then see how we can ensure that the manifold shares them.

2.1.1 Topology of the real line $\mathbb{R}$ and of $\mathbb{R}^d$

The topological properties we are interested in are expressed in terms of 'open sets', which are defined in the following way. An *open interval* (a, b) is the set of all points (real numbers) x such that $a < x < b$:



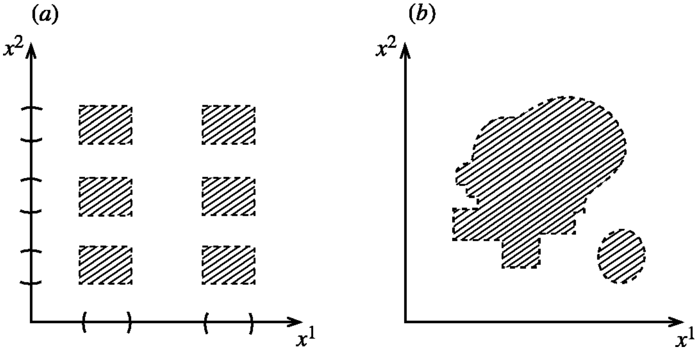
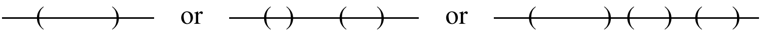


Figure 2.3. (a) An open set in $\mathbb{R}^2$. It is a union of open rectangles constructed from unions of open intervals in the two copies of $\mathbb{R}$ which form the x^1 and x^2 axes. (b) Another open set in $\mathbb{R}^2$, which can be constructed as a union of open rectangles.

The end points $x = a$ and $x = b$ are excluded. Consequently, *any* point x in (a, b) can be surrounded by another open interval $(x - \epsilon, x + \epsilon)$, all of whose points are also in (a, b). For example, however close x is to a, it cannot be equal to a. There are always points between a and x, and if x is closer to a than to b, we can take $\epsilon = (x - a)/2$. An *open set* of $\mathbb{R}$ is defined as any union of 1, 2, 3, ... open intervals:



etc. (The *union* $A \cup B \cup C \cdots$ of a number of sets is defined as the set of all points that belong to at least one of $A, B, C, \ldots$. The *intersection* $A \cap B \cap C \cdots$ is the set of all points that belong to all the sets $A, B, C, \ldots$.) In addition, the empty set, which contains no points, is defined to be an open set.

The space $\mathbb{R}^2$ is the set of all pairs of real numbers (x^1, x^2), which can be envisaged as an infinite plane. The definition of open sets is easily extended to $\mathbb{R}^2$, as illustrated in figure 2.3. If x^1 lies in a chosen open interval on the horizontal axis, and x^2 in a chosen open interval on the vertical axis, then (x^1, x^2) lies in an open rectangle corresponding to these two intervals. Any union of open rectangles is an open set. Since the rectangles can be arbitrarily small, we can say that any region bounded by a closed curve, but excluding points actually on the curve, is also an open set, and so is any union of such regions. Obviously, the same ideas can be further extended to $\mathbb{R}^d$, which is the set of all d-tuples of real numbers $(x^1, x^2, \ldots, x^d)$.

An important use of open sets is to define continuous functions. Consider, for instance, a function f which takes real numbers x as arguments and has real-number values $y = f(x)$. An example is shown in figure 2.4. The *inverse image* of a set of points on the y axis is the set of all those points on the x axis for which $f(x)$ belongs to the original set. Then we say that f is continuous if the

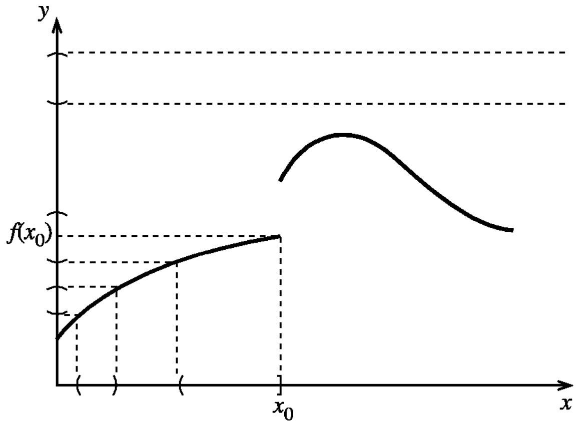


Figure 2.4. The graph $y = f(x)$ of a function which is discontinuous at x_0. Any open interval of y which includes $f(x_0)$ has an inverse image on the x axis which is not open. The inverse image of an interval in y which contains no values of $f(x)$ is the empty set.

inverse image of any open set on the y axis is an open set on the x axis. The example shown fails to be continuous because the inverse image of any open interval containing $f(x_0)$ contains an interval of the type $(x_1, x_0]$, which includes the end point x_0 and is therefore not open. (Readers who are not at home with this style of argument should spend a short while considering the implications of these definitions: why, for example, is it necessary to include not only open intervals but also their unions and the empty set as open sets?)

The open sets of $\mathbb{R}^d$ have two fairly obvious properties: (i) any union of open sets is itself an open set; (ii) any intersection of a finite number of open sets is itself an open set. Given any space (by which we mean a set of points), suppose that a collection of subsets of its points is specified, such that any union or finite intersection of them also belongs to the collection. We also specify that the entire space (which counts as a subset of itself) and the empty set belong to the collection. Then the subsets in this collection may, by analogy, be called *open sets*. The collection of open sets is called a *topology* and the space, together with its topology, is called a *topological space*. It is, of course, possible to endow a given space with many different topologies. For example, the collection of all subsets of the space clearly satisfies all the above conditions, and is called the *discrete topology*. By endowing the real line with this topology, we would obtain a new definition of continuity—it would not be a particularly useful definition, however, as any function at all would turn out to be continuous. The particular topology of $\mathbb{R}^d$ described above is called the *natural topology* and is the one we shall always use.

It is important to realize that a topology is quite independent of any notion of distance. For instance, a sheet of paper may be regarded as a part of $\mathbb{R}^2$.

The natural topology reflects the way in which its points fit together to form a coherent structure. If it is used to draw figures in Euclidean geometry, then the distance D between two points is defined by the Pythagoras rule as $D = \left[(\Delta x)^2 + (\Delta y)^2\right]^{1/2}$. But it might equally well be used to plot the mean atmospheric concentration of carbon monoxide in central London (represented by y) as a function of time (represented by x), in which case D would have no sensible meaning.

A topology imposes two kinds of structure on the space. The *local topology*—the way in which open sets fit inside one another over small regions—determines the way in which notions like continuity apply to the space. The *global topology*—the way in which the open sets can be made to cover the whole space—determines its overall structure. Thus, the plane, sphere and torus have the same local structure but different global structures. Physically, we have no definite information about the global topology of spacetime, but its local structure seems to be very similar to that of $\mathbb{R}^4$ (though we shall encounter speculative theories that call this apparently simple observation into question).

2.1.2 Differentiable spacetime manifold

In order that our model of space and time should be able to support continuous and differentiable functions of the sort that we rely on to do physics, we want it (for now) to have the same local topology as $\mathbb{R}^4$. First of all, then, it must be a topological space. That is, it must have a collection of open sets, in terms of which continuous functions can be defined. Second, the structure of these open sets must be similar, within small regions, to the natural topology of $\mathbb{R}^4$. To this end, we demand that every point of the space belong to at least one open set, all of whose points can be put into a one-to one correspondence with the points of some open set of $\mathbb{R}^4$. More technically, the correspondence is a one-to-one mapping of the open set of the space *onto* the open set of $\mathbb{R}^4$, which is to say that every point of the open set in the space has a unique image point in the open set of $\mathbb{R}^4$ and *vice versa*. We further demand that this mapping be continuous, according to our previous definition. When these conditions are met, the space is called a *manifold*. The existence of continuous mappings between the manifold and $\mathbb{R}^4$ implies that a function f defined on the manifold (that is, one that has a value $f(P)$ for each point P of the manifold) can be re-expressed as a function g defined on $\mathbb{R}^4$, so that $f(P) = g(x^0, \ldots, x^3)$, where $(x^0, \ldots, x^3)$ is the point of $\mathbb{R}^4$ corresponding to P. In this way, continuous functions defined on the manifold inherit the characteristics of those defined on $\mathbb{R}^4$.

This definition amounts to saying that the manifold can be covered by patches, in each of which a four-dimensional coordinate system can be set up, as illustrated in figure 2.5 for the more easily drawn case of a two-dimensional manifold. Normally, of course, many different coordinate systems can be set up on any part of the manifold. The definition also ensures that, within the range of coordinate values corresponding to a given patch, there exists a point of the

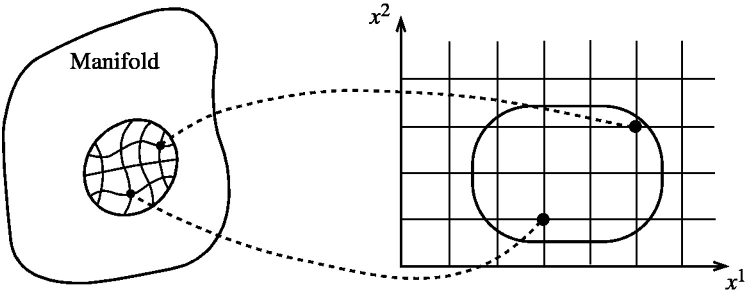


Figure 2.5. A coordinate patch on a two-dimensional manifold. Each point in the patch is mapped to a unique image point in a region of $\mathbb{R}^2$ and *vice versa*.

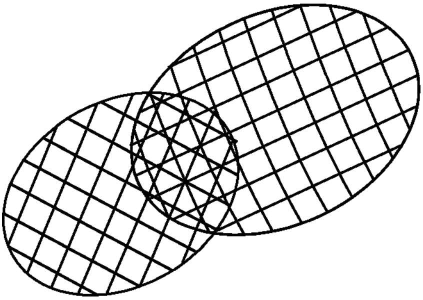


Figure 2.6. Two overlapping coordinate patches. A point in the overlap region can be identified using either set of coordinates.

manifold for each set of coordinate values—so there are no points ‘missing’ from the manifold, and also that there are no ‘extra’ points that cannot be assigned coordinates. Within a coordinate patch, a quantity such as an electric potential, which has a value at each point of the manifold, can be expressed as an ordinary function of the coordinates of the point. Often, we shall expect such functions to be *differentiable* (that is, to possess unique partial derivatives with respect to each coordinate at each point of the patch).

Suppose we have two patches, each with its own coordinate system, that partly or wholly overlap, as in figure 2.6. Each point in the overlap region has two sets of coordinates, say $(x^0, \dots, x^3)$ and $(y^0, \dots, y^3)$, and the y coordinates can be expressed as functions of the x coordinates: $y^0 = y^0(x^0, \dots, x^3)$, etc. Given ‘reasonable’ coordinate systems, we might suppose that a function which is differentiable when expressed in terms of the x^μ ought also to be differentiable when expressed in terms of the y^μ. This will indeed be true if the transformation functions $y^\mu(x)$ are differentiable. If the manifold can be completely covered by a set of coordinate patches, in such a way that all of these transformation functions are differentiable, then we have a *differentiable manifold*. In order for a function

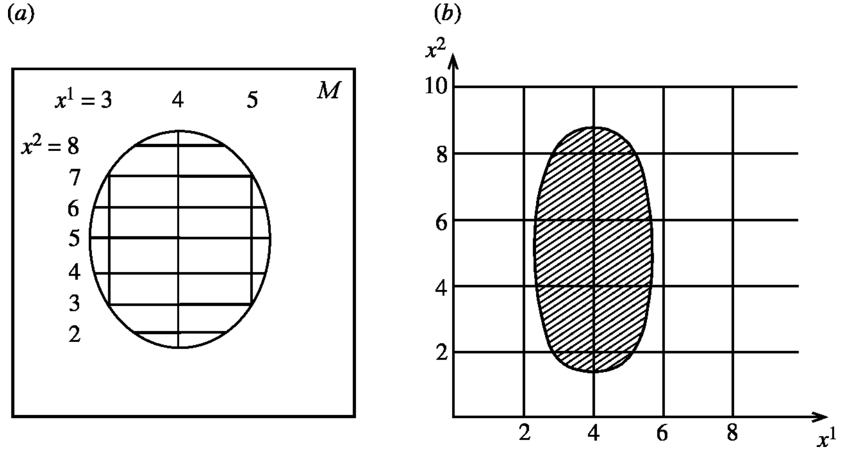


Figure 2.7. (a) A manifold M, part of the surface of this page, with a coordinate patch. (b) Part of $\mathbb{R}^2$, showing the coordinate values used in (a).

to remain differentiable at least n times after a change of coordinates, at least the first n derivatives of all the transformation functions must exist. If they do, then we have what is called a C^n manifold. Intuitively, we might think it possible to define functions of space and time that can be differentiated any number of times, for which we would need $n = \infty$. We shall indeed take a C^∞ manifold as the basis for our model spacetime. Mathematically, though, this is a rather strong assumption, and for many physical purposes it would be sufficient to take, say, $n = 4$.

2.1.3 Summary and examples

Our starting point for a model of space and time is a C^∞ manifold. The essence of the technical definition described above is, first, that it is possible to set up a local coordinate system covering any sufficiently 'small' region and, second, that it is possible to define functions on the manifold that are continuous and differentiable in the usual sense. It is, of course, perfectly possible to define functions that are neither continuous nor differentiable. The point is that, if a function fails to be continuous or differentiable, this will be the fault either of the function itself or of our choice of coordinates, but not the fault of the manifold. The word 'small' appears in inverted commas because, as I have emphasized, there is as yet no definite notion of length: it simply means that it may well not be possible to cover the entire manifold with a single coordinate system. The coordinate systems themselves are not part of the structure of the manifold. They serve merely as an aid to thought, providing a practical means of specifying properties of sets of points belonging to the manifold.

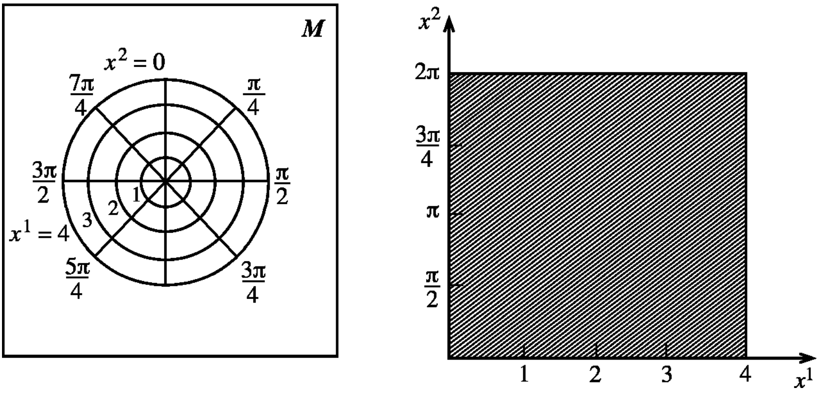


Figure 2.8. Same as figure 2.7, but using different coordinates.

The following examples illustrate, in terms of two-dimensional manifolds, some of the important ideas. Figure 2.7(*a*) shows a manifold, M, which is part of the surface of the paper on which it is printed. For the sake of argument, I am asking readers to suppose that this surface is perfectly smooth, rather than being composed of tiny fibres. For the definitions to work, we must take the manifold to be the interior of the rectangular region, excluding points *on* the boundary. The interior of the roughly circular region is a coordinate patch. Inside it are drawn some of the grid lines by means of which we assign coordinates x^1 and x^2 to each point. Figure 2.7(*b*) is a pictorial representation of part of the space $\mathbb{R}^2$ of pairs of coordinates. The interior of the shaded region represents the coordinates actually used. To every point of this region there corresponds a point of the coordinate patch in M and *vice versa*. Figure 2.8 shows a similar arrangement, using a different coordinate system. Here, again, the *interior* of the shaded region of $\mathbb{R}^2$ represents the open set of points that correspond uniquely to points of the coordinate patch. As before, the boundary of the coordinate patch and the corresponding line $x^1 = 4$ in $\mathbb{R}^2$ are excluded. Also excluded, however, are the boundary lines $x^1 = 0$, $x^2 = 0$ and $x^2 = 2\pi$ in $\mathbb{R}^2$, which means that points on the line labelled by $x^2 = 0$ in M do not, in fact, belong to the coordinate patch. Since the coordinate system is obviously usable, even when these points are included, their exclusion may seem like an annoying piece of bureaucracy: however, it is essential to apply the rules correctly if the definitions of continuity and differentiability are to work smoothly. For example, the function $g(x^1, x^2) = x^2$ is continuous throughout $\mathbb{R}^2$, but the corresponding function on M is discontinuous at $x^2 = 0$.

It should be clear that, whereas a single coordinate patch like that in figure 2.7 can be extended to cover the whole of M, at least two patches of the kind shown in figure 2.8 would be needed. Readers should also be able to convince themselves that, if M were the two-dimensional surface of a sphere, no single patch of any kind could cover all of it. These examples also illustrate the fact that, although the coordinates which label the points of M have definite numerical values, these values do not, in themselves, supply any notion of a distance between two points. The distance along some curve in M *may* be defined by some suitable rule, such as (i) 'use a ruler' or (ii) 'measure the volume of ink used by a standard pen to trace the curve' or, given a particular coordinate system, (iii) 'use the mathematical expression $D =$ (function of coordinates)'. Any such rule imposes an additional structure—called a *metric*—which is not inherent in the manifold. In particular, there is no naturally occurring function for use in (iii). Any specific function, such as the Pythagoras expression, would have quite different effects when applied to different coordinate systems, and the definition of the manifold certainly does not single out a special coordinate system to which that function would apply. We do have a more or less unambiguous means of determining distances on a sheet of paper, and this is because the paper, in addition to the topological properties it possesses as a manifold, has physical properties that enable us to apply a definite measuring procedure. The same is true of space and time and, although we have made some initial assumptions about their topological structure, we have yet to find out what physical properties determine their metrical structure.

2.2 Tensors

From our discussion so far, it is apparent that coordinate systems can be dangerous, even though they are often indispensable for giving concrete descriptions of a physical system. We have seen that the topology of a manifold such as that of space and time may permit the use of a particular coordinate system only within a small patch. Suppose, for the sake of argument, that the surface of the Earth is a smooth sphere. We encounter no difficulty in drawing, say, the street plan of a city on a flat sheet of paper using Cartesian coordinates, but we should obviously be misled if we assumed that this map could be extended straightforwardly to cover the whole globe. By assuming that two-dimensional Euclidean geometry was valid on the surface of the Earth, we should be making a mistake, owing to the curvature of the spherical surface, but the mistake would not become apparent as long as we made measurements only within a region the size of a city. Likewise, physicists before Einstein assumed that a frame of reference fixed on the Earth would be inertial, except for effects of the known orbital motion of the Earth around the Sun and its rotation about its own axis, which could be corrected for if necessary. According to Einstein, however, this assumption is also mistaken. It fails to take account of the true geometry of space and time in much the same way that, by treating a city plan as a Euclidean plane, we fail to take account of the true geometry of the Earth. The mistake only becomes apparent, however, when we make precise observations of gravitational phenomena.

The difficulty here is that we often express the laws of physics in the form which, we believe, applies to inertial frames. If we do not know, *a priori*, what the true geometry of space and time is, then we do not know whether any given frame is truly inertial. Therefore, we need to express our laws in a way that does not rely on our making any special assumption about the coordinate system. There are two ways of achieving this. The method adopted by Einstein himself is to write our equations in a form that applies to *any* coordinate system: the mathematical techniques for doing this constitute what is called *tensor analysis*. The other, more recent method is to write them in a manner that makes no reference to coordinate systems at all: this requires the techniques of *differential geometry*. For our purposes, these two approaches are entirely equivalent, but each has its own advantages and disadvantages in terms of conceptual and notational clarity. So far as I can, I will follow a middle course, which seems to me to maximize the advantages. Both techniques deal with objects called *tensors*. Tensor analysis, like elementary vector analysis, treats them as being defined by sets of components, referred to particular coordinate systems. Differential geometry treats them as entities in their own right, which may be described in terms of components, but need not be. When components are used, the two techniques become identical, so there is no difficulty in changing from one description to the other.

Many, though not all, of the physical objects that inhabit the spacetime manifold will be described by tensors. A *tensor* at a point P of the manifold refers only to that point. A *tensor field* assigns some property to every point of the manifold, and most physical quantities will be described by tensor fields. (For brevity, I shall often follow custom by referring to a tensor field simply as a 'tensor', when the meaning is obvious from the context.) Tensors and tensor fields are classified by their *rank*, a pair of numbers $\binom{a}{b}$.

Rank $\binom{0}{0}$ tensors, also called *scalars*, are simply real numbers. A *scalar field* is a real-valued function, say $f(P)$, which assigns a real number to each point of the manifold. If our manifold were just the three-dimensional space encountered in Newtonian physics, then at a particular instant in time, an electric potential $V(P)$ or the density of a fluid $\rho(P)$ would be examples of scalar fields. In relativistic physics, these and all other simple examples I can think of are not true scalars, because their definitions depend in one way or another on the use of specific coordinate systems or on metrical properties of the space that our manifold does not yet possess. For the time being, however, no great harm will be done if readers bear these examples in mind. If we introduce coordinates x^μ, then we can express $f(P)$ as an algebraic function $f(x^\mu)$. (For convenience, I am using the same symbol f to denote two different, though related functions: we have $f(x^\mu) = f(P)$ when x^μ are the coordinates of the point P.) In a different coordinate system, where P has the coordinates $x^{\mu'}$, the same quantity will be described by a new algebraic function $f'(x^{\mu'})$, related to the old one by

$$f'(x^{\mu'}) = f(x^\mu) = f(P). \tag{2.9}$$

In tensor analysis, this transformation law is taken to *define* what is meant by a scalar field.

Rank $\binom{1}{0}$ tensors are called *vectors* in differential geometry. They correspond to what are called *contravariant vectors* in tensor analysis. The prototypical vector is the tangent vector to a curve. In ordinary Euclidean geometry, the equation of a curve may be expressed parametrically by giving three functions $x(\lambda)$, $y(\lambda)$ and $z(\lambda)$, so that each point of the curve is labelled by a value of λ and the functions give its coordinates. If λ is chosen to be the distance along the curve from a given starting point, then the tangent vector to the curve at the point labelled by λ has components $(\mathrm{d}x/\mathrm{d}\lambda, \mathrm{d}y/\mathrm{d}\lambda, \mathrm{d}z/\mathrm{d}\lambda)$. In our manifold, we have not yet given any meaning to 'distance along the curve', and we want to avoid defining vectors in terms of their components relative to a specific coordinate system. Differential geometry provides the following indirect method of generalizing the notion of a vector to any manifold. Consider, in Euclidean space, a differentiable function $f(x, y, z)$. This function has, in particular, a value $f(\lambda)$ at each point of the curve, which we obtain by substituting for x, y and z the appropriate functions of λ. The rate of change of f with respect to λ is

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \frac{\mathrm{d}x}{\mathrm{d}\lambda}\frac{\partial f}{\partial x} + \frac{\mathrm{d}y}{\mathrm{d}\lambda}\frac{\partial f}{\partial y} + \frac{\mathrm{d}z}{\mathrm{d}\lambda}\frac{\partial f}{\partial z} \tag{2.10}$$

so, by choosing $f = x$, $f = y$ or $f = z$, we can recover from this expression each component of the tangent vector. All the information about the tangent vector is contained in the differential operator $\mathrm{d}/\mathrm{d}\lambda$, and in differential geometry this operator is defined to *be* the tangent vector.

A little care is required when applying this definition to our manifold. We can certainly draw a continuous curve on the manifold and label its points continuously by a parameter λ. What we cannot yet do is select a special parameter that measures distance along it. Clearly, by choosing different parametrizations of the curve, we shall arrive at different definitions of its tangent vectors. It is convenient to refer to the one-dimensional set of points in the manifold as a *path*. Then each path may be parametrized in many different ways, and we regard each parametrization as a distinct *curve*. This has the advantage that each curve, with its parameter λ, has a unique tangent vector $\mathrm{d}/\mathrm{d}\lambda$ at every point. Suppose we have two curves, corresponding to the same path, but with parameters λ and μ that are related by $\mu = a\lambda + b$, a and b being constants. The difference is obviously a rather trivial one and the two parameters are said to be *affinely related*.

If we now introduce a coordinate system, we can resolve a vector into components, in much the same way as in Euclidean geometry. At this point, it is useful to introduce two abbreviations into our notation. First, we use the symbol ∂_μ to denote the partial derivative $\partial/\partial x^\mu$. Second, we shall use the *summation convention*, according to which, if an index such as μ appears in an expression twice, once in the upper position and once in the lower position, then a sum over the values $\mu = 0 \ldots 3$ is implied. (More generally, in a d-dimensional manifold, the sum is over the values $0 \ldots (d-1)$. In contexts other than spacetime geometry, there may be no useful distinction between upper and lower indices, and repeated indices implying a sum may both appear in the same position.) I shall use bold capital letters to denote vectors, such as $\boldsymbol{V} = \mathrm{d}/\mathrm{d}\lambda$. If, then, a curve is represented in a particular coordinate system by the functions $x^\mu(\lambda)$, we can write

$$\boldsymbol{V} \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda} = \sum_{\mu=0}^{3} \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\partial}{\partial x^\mu} \equiv V^\mu \partial_\mu \equiv V^\mu \boldsymbol{X}_\mu \tag{2.11}$$

where the partial derivatives $\boldsymbol{X}_\mu = \partial/\partial x^\mu$ are identified as the basis vectors in this system and V^μ are the corresponding components of $\boldsymbol{V}$. Note that components of a vector are labelled by upper indices and basis vectors by lower ones. In a new coordinate system, with coordinates $x^{\mu'}$, and basis vectors $\boldsymbol{X}_{\mu'} = \partial/\partial x^{\mu'}$, the chain rule $\partial_\mu = (\partial x^{\mu'}/\partial x^\mu)\partial_{\mu'}$ shows that the same vector has components

$$V^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\mu} V^\mu . \tag{2.12}$$

In tensor analysis, a contravariant vector is defined by specifying its components in some chosen coordinate system and requiring its components in any other system to be those given by the transformation law (2.12). It will be convenient to denote the transformation matrix by

$$\Lambda^{\mu'}{}_{\mu} = \frac{\partial x^{\mu'}}{\partial x^\mu} . \tag{2.13}$$

The convention of placing a prime on the index μ' to indicate that x^μ and $x^{\mu'}$ belong to different coordinate systems, rather than writing, say, x'^μ, is useful here in indicating to which system each index on Λ refers. Using the chain rule again, we find

$$\Lambda^{\mu}{}_{\nu'}\Lambda^{\nu'}{}_{\sigma} = \frac{\partial x^\mu}{\partial x^{\nu'}}\frac{\partial x^{\nu'}}{\partial x^\sigma} = \frac{\partial x^\mu}{\partial x^\sigma} = \delta^\mu_\sigma \tag{2.14}$$

so the matrix $\Lambda^{\mu}{}_{\nu'}$ is the inverse of the matrix $\Lambda^{\nu'}{}_{\mu}$.

Rank $\binom{0}{1}$ tensors are called *one-forms* in differential geometry or *covariant vectors* in tensor analysis. Consider the scalar product $\boldsymbol{u} \cdot \boldsymbol{v}$ of two Euclidean vectors. Normally, we regard this product as a rule that combines two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to produce a real number. As we shall see, this scalar product involves metrical properties of Euclidean space that our manifold does not yet possess. There is, however, a different point of view that can be transferred to manifold. For a given vector $\boldsymbol{u}$, the symbol $\boldsymbol{u}\cdot$ can be regarded as defining a *function*, whose argument is a vector, say $\boldsymbol{v}$, and whose value is the real number $\boldsymbol{u} \cdot \boldsymbol{v}$. The function $\boldsymbol{u}\cdot$ is linear. That is to say, if we give it the argument $a\boldsymbol{v} + b\boldsymbol{w}$, where $\boldsymbol{v}$ and $\boldsymbol{w}$ are any two vectors, and a and b are any two real numbers, then $\boldsymbol{u} \cdot (a\boldsymbol{v} + b\boldsymbol{w}) = a\boldsymbol{u} \cdot \boldsymbol{v} + b\boldsymbol{u} \cdot \boldsymbol{w}$. This is, in fact, the definition of a one-form. In our manifold, a one-form, say ω, is a real-valued, linear function whose argument is a vector: $\omega(\boldsymbol{V}) =$ (real number). Because the one-form is a linear function, its value must be a linear combination of the components of the vector:

$$\omega(\boldsymbol{V}) = \omega_\mu V^\mu . \tag{2.15}$$

The coefficients ω_μ are the components of the one-form, relative to the coordinate system in which $\boldsymbol{V}$ has components V^μ. A *one-form field* is defined in the same way as a linear function of vector fields, whose value is a scalar field. In the definition of linearity, a and b may be any two scalar fields.

The expression (2.15) is, of course, similar to the rule for calculating the scalar product of two Euclidean vectors from their components. Nevertheless, it is clear from their definitions that vectors and one-forms are quite different things, and (2.15) does not allow us to form a scalar product of two vectors.

An example of a one-form field is the gradient of a scalar field f, whose components are $\partial_\mu f$. Notice the consistency of the convention for placing indices: the components of a one-form have indices that naturally appear in the lower position. Call this gradient one-form ω_f. If $\boldsymbol{V} = \mathrm{d}/\mathrm{d}\lambda$ is the tangent vector to a curve $x^\mu(\lambda)$, then the new scalar field $\omega_f(\boldsymbol{V})$ is the rate of change of f along the curve:

$$\omega_f(\boldsymbol{V}) = \frac{\partial f}{\partial x^\mu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}f}{\mathrm{d}\lambda}. \tag{2.16}$$

Since vectors and one-forms exist independently of any coordinate system, the function $\omega(\boldsymbol{V})$ given in (2.15) must be a true scalar field—it must have the same value in any coordinate system. This means that the matrix which transforms the components of a one-form between two coordinate systems must be the inverse of that which transforms the components of a vector:

$$\omega_{\mu'} = \omega_\mu \Lambda^\mu{}_{\mu'} = \omega_\mu \frac{\partial x^\mu}{\partial x^{\mu'}}. \tag{2.17}$$

Then, on transforming (2.15), we get

$$\omega(\boldsymbol{V}) = \omega_{\mu'} V^{\mu'} = \omega_\mu \Lambda^\mu{}_{\mu'} \Lambda^{\mu'}{}_\nu V^\nu = \omega_\mu \delta^\mu{}_\nu V^\nu = \omega_\mu V^\mu . \tag{2.18}$$

In tensor analysis, a covariant vector is defined by requiring that its components obey the transformation law (2.17). Clearly, this is indeed the correct way of transforming a gradient.

Rank $\binom{a}{b}$ tensors and tensor fields can be defined in a coordinate-independent way, making use of the foregoing definitions of vectors and one-forms, and I shall say more about this in §3.7. For our present purposes, however, it becomes rather easier at this point to adopt the tensor analysis approach of defining higher-rank tensors in terms of their components. A tensor of *contravariant rank* a and *covariant rank* b has, in a d-dimensional manifold, d^{a+b} components, labelled by a upper indices and b lower ones. The tensor may be specified by giving all of its components relative to some chosen coordinate system. In any other system, the components are then given by a transformation law that generalizes those for vectors and one-forms in an obvious way:

$$T^{\alpha'\beta'\dots}{}_{\mu'\nu'\dots} = \Lambda^{\alpha'}{}_{\alpha}\Lambda^{\beta'}{}_{\beta}\cdots\Lambda^{\mu}{}_{\mu'}\Lambda^{\nu}{}_{\nu'}\cdots T^{\alpha\beta\dots}{}_{\mu\nu\dots}. \qquad (2.19)$$

From this we can see how to construct laws of physics in a way that will make them true in any coordinate system. Suppose that a fact about some physical system is expressed in the form $S = T$, where S and T are tensors of the same rank. On multiplying this equation on both sides by the appropriate product of Λ matrices, we obtain the equation $S' = T'$, which expresses the same fact, in an equation of the same form, but now applies to the new coordinate system. The point that may require some effort is to make sure that S and T really *are* tensors that transform in the appropriate way.

If ω is a one-form and V a vector, then the d^2 quantities $T^{\nu}_{\mu} = \omega_{\mu}V^{\nu}$ are the components of a rank $\binom{1}{1}$ tensor. As we saw in (2.15), by setting $\mu = \nu$ and carrying out the implied sum, we obtain a single number, which is a scalar (or a rank $\binom{0}{0}$ tensor). This process is called *contraction*. Given any tensor of rank $\binom{a}{b}$, with $a \geq 1$ and $b \geq 1$, we may contract an upper index with a lower one to obtain a new tensor of rank $\binom{a-1}{b-1}$. Readers should find it an easy matter to check from (2.19) that, for example, the object $S^{\alpha\gamma\dots}{}_{\nu\dots} = T^{\alpha\beta\gamma\dots}{}_{\beta\nu\dots}$ does indeed transform in the right way.

2.3 Extra Geometrical Structures

Two geometrical structures are needed to endow our manifold with the familiar properties of space and time: (i) the notion of *parallelism* is represented mathematically by an *affine connection*; (ii) the notions of *length* and *angle* are represented by a *metric*. In principle, these two structures are quite independent. In Euclidean geometry, of course, it is perfectly possible to define what we mean by parallel lines in terms of distances and angles, and this is also true of the structures that are most commonly used in general-relativistic geometry. Thus there is, as we shall see, a special kind of affine connection that can be deduced from a metric. It is called a *metric connection* (or sometimes, the *Levi-Civita connection*). We shall eventually assume that the actual geometry of space and time is indeed described by a metric connection. From a theoretical point of view, however, it is instructive to understand the distinction between those geometrical ideas that rely only on an affine connection and those that require a metric. Moreover, there are manifolds other than spacetime that play important roles in physics (in particular, those connected with the gauge theories of particle physics), which possess connections, but do not necessarily possess metrics. To emphasize this point, therefore, I shall deal first with the affine connection, then with the metric, and finally with the metric connection.

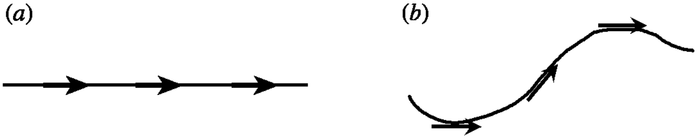


Figure 2.9. (*a*) A geodesic curve: successive tangent vectors are parallel to each other. (*b*) A non-geodesic curve: successive tangent vectors are not parallel.

2.3.1 The affine connection

There are four important geometrical tools provided by an affine connection: the notion of *parallelism*, the notion of *curvature*, the *covariant derivative* and the *geodesic*. Let us first understand what it is good for.

a) Newton's first law of motion claims that 'a body moves at constant speed in a straight line unless it is acted on by a force'. In general relativity, we shall replace this with the assertion that 'a test particle follows a geodesic curve unless it is acted on by a non-gravitational force'. As we saw earlier, gravitational forces are going to be interpreted in terms of spacetime geometry, which itself is modified by the presence of gravitating bodies. By a 'test particle', we mean one that responds to this geometry, but does not modify it significantly. A *geodesic* is a generalization of the straight line of Euclidean geometry. It is defined, roughly, as a curve whose tangent vectors at successive points are parallel, as illustrated in figure 2.9. Given a definition of 'parallel', as provided by the connection, this is perhaps intuitively recognizable as the natural state of motion for a particle that is not disturbed by external influences.

b) The equations of physics, which we wish to express entirely in terms of tensors, frequently involve the derivatives of vector or tensor fields. Now, the derivatives of a scalar field $\partial_\mu f$ are, as we have seen, the components of a one-form field. However, the derivatives of the components of a vector field, $\partial_\mu V^\nu$, are not the components of a tensor field, even though they are labelled by a contravariant and a covariant index. On transforming these derivatives to a new coordinate system, we find

$$\begin{aligned}\partial_{\mu'} V^{\nu'} &= \Lambda^{\mu}{}_{\mu'} \partial_\mu (\Lambda^{\nu'}{}_{\nu} V^\nu)\\ &= \Lambda^{\mu}{}_{\mu'}\Lambda^{\nu'}{}_{\nu}\partial_\mu V^\nu + \Lambda^{\mu}{}_{\mu'}(\partial_\mu \Lambda^{\nu'}{}_{\nu})V^\nu. \qquad (2.20)\end{aligned}$$

Because of the last term, this does not agree with the transformation law for a second-rank tensor. The affine connection will enable us to define what is called a *covariant derivative*, ∇_μ, whose action on a vector field is of the form $\nabla_\mu V^\nu = \partial_\mu V^\nu + \text{(connection term)}$. The transformation of the extra term involving the affine connection will serve to cancel the unwanted part in (2.20), so that $\nabla_\mu V^\nu$ will be a tensor.

c) The fact that the functions $\partial_\mu V^\nu$ do not transform as the components of a tensor indicates that they have no coordinate-independent meaning. To see what

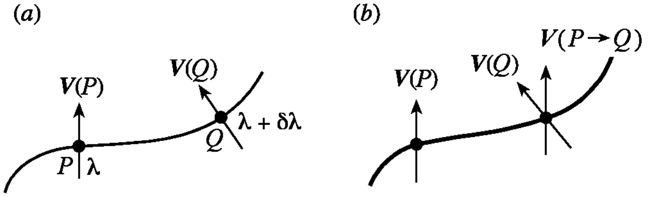


Figure 2.10. $\boldsymbol{V}(P)$ and $\boldsymbol{V}(Q)$ are the vectors at P and Q belonging to the vector field $\boldsymbol{V}$. $\boldsymbol{V}(P \to Q)$ is the vector at Q which results from parallelly transporting $\boldsymbol{V}(P)$ along the curve.

goes wrong, consider the derivative of a component of a vector field along a curve, as illustrated in figure 2.10(*a*), where P and Q are points on the curve with parameters λ and $\lambda + \delta\lambda$ respectively. The derivative at P is

$$\frac{\mathrm{d}V^{\mu}}{\mathrm{d}\lambda} = \frac{\mathrm{d}x^{\nu}}{\mathrm{d}\lambda}\frac{\partial V^{\mu}}{\partial x^{\nu}} = \lim_{\delta\lambda\to 0}\frac{V^{\mu}(Q) - V^{\mu}(P)}{\delta\lambda}. \tag{2.21}$$

For a scalar field, which has unique values at P and Q, such a derivative makes good sense. However, the values at P and Q of the components of a vector field depend on the coordinate system to which they are referred. It is easy to make a change of coordinates such that, for example, $V^{\mu}(Q)$ is changed while $V^{\mu}(P)$ is not, and so the difference of these two quantities has no coordinate-independent meaning. If we try to find the derivative of the vector field itself, we shall encounter the expression $\boldsymbol{V}(Q) - \boldsymbol{V}(P)$. Now, $\boldsymbol{V}(P)$ is the tangent vector to some curve passing through P (though not necessarily the one shown in figure 2.10(*a*)) and $\boldsymbol{V}(Q)$ is the tangent vector to some curve passing through Q. The difference of two vectors at P is another vector at P: each vector is tangent to some curve passing through P. However, $\boldsymbol{V}(Q) - \boldsymbol{V}(P)$ is not, in general, the tangent vector to a curve at a specific point. It is not, therefore, a vector and has, indeed, no obvious significance at all.

To define a meaningful derivative of a vector field, we need to compare two vectors at the same point, say Q. Therefore, we construct a new vector $\boldsymbol{V}(P \to Q)$, which exists at Q but represents $\boldsymbol{V}(P)$. Then a new vector, $\mathrm{D}\boldsymbol{V}/\mathrm{d}\lambda$, which will be regarded as the derivative of $\boldsymbol{V}$ along the curve, may be defined as

$$\left.\frac{\mathrm{D}\boldsymbol{V}}{\mathrm{d}\lambda}\right|_{P} = \lim_{\delta\lambda\to 0}\frac{\boldsymbol{V}(Q) - \boldsymbol{V}(P \to Q)}{\delta\lambda}. \tag{2.22}$$

In the limit, of course, Q coincides with P and this is where the new vector exists. There is no natural way in which a vector at Q corresponds to a vector at P, so we must provide a rule to define $\boldsymbol{V}(P \to Q)$ in terms of $\boldsymbol{V}(P)$. This rule is the affine connection. In figure 2.10(*b*), $\boldsymbol{V}(P \to Q)$ is shown as a vector at Q that is parallel to $\boldsymbol{V}(P)$. The figure looks this way because of the Euclidean properties of the paper on which it is printed. Mathematically, the affine

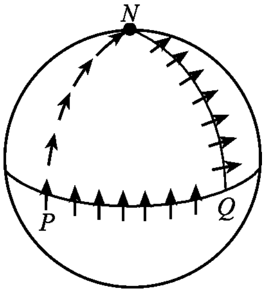


Figure 2.11. Parallel transport of a vector from P to Q on a spherical surface by two routes.

connection *defines* what it means for a vector at Q to be parallel to one at P: it is said to define *parallel transport* of a vector along the curve. From a mathematical point of view, we are free to specify the affine connection in any way we choose. Physically, on the other hand, we shall need to find out what the affine connection is, with which nature has actually provided us, and we shall address this problem in due course. It might be thought that a vector which represents $\boldsymbol{V}(P)$ should not only be parallel to it but also have the same length. In Euclidean geometry, the magnitude of a vector is $(\boldsymbol{v} \cdot \boldsymbol{v})^{1/2}$ and, as we have seen, the scalar product needs a metric for its definition. The metric connection, mentioned above, does indeed define parallel transport in a manner that preserves the magnitude of the transported vector.

The concrete definition of parallel transport is most clearly written down by choosing a coordinate system. If P and Q lie on a curve $x^\mu(\lambda)$ and are separated by an infinitesimal parameter distance $\delta\lambda$, then the components of $\boldsymbol{V}(P \to Q)$ are defined by

$$V^\mu(P \to Q) = V^\mu(P) - \delta\lambda\Gamma^\mu{}_{\nu\sigma}(P)V^\nu(P)\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} \tag{2.23}$$

and the functions $\Gamma^\mu{}_{\nu\sigma}$ are called the *affine connection coefficients*. These coefficients exist at each point of the manifold and are not associated with any particular curve. However, the rule (2.23) for parallel transport involves, in addition to the vector $\boldsymbol{V}$ itself, both the connection coefficients and the tangent vector $\mathrm{d}x^\sigma/\mathrm{d}\lambda$, so parallel transport is defined only along a curve. To transport $\boldsymbol{V}$ along a curve by a finite parameter distance, we have to integrate (2.23). If we wish to transport a vector from an initial point P to a final point Q, we must choose a curve, passing through both P and Q, along which to transport it. There will usually be many such curves and it is vital to realize that the vector which finally arrives at Q depends on the route taken: the functions $\Gamma^\mu{}_{\nu\sigma}$ will generally not take the same values along two different curves. This fact lies at the root of the idea of the *curvature* of a manifold, as we shall see shortly.

The idea of parallel transport is illustrated in figure 2.11, which shows the surface of a Euclidean sphere. For the purposes of this example, we assume the usual metrical properties of Euclidean space, so that distances and angles have their usual meanings. The manifold we consider is the two-dimensional surface of the sphere, so every vector is tangential to this surface. P and Q are points on the equator, separated by a quarter of its circumference, and N is the north pole. The equator and the curves PN and QN are parts of great circles on the sphere and are 'straight lines' as far as geometry on the spherical surface is concerned: one would follow such a path by walking straight ahead on the surface of a perfectly smooth Earth. Consider a vector $\boldsymbol{V}(P)$ that points due north—it is a tangent vector at P to the curve PN. We shall transport this vector to Q, first along the equator and second via the north pole. The rule for parallel transport of a vector along a straight line is particularly simple: the angle between the vector and the line remains constant. For transport along the equator, the vector clearly points north at each step and so $\boldsymbol{V}(P \to Q)$ also points north along QN. Along PN, the vector also points north, so on arrival at the pole it is perpendicular to QN. On its way south, it stays perpendicular to QN. Thus, the transported vector $\boldsymbol{V}(P \to Q)$ as defined by the polar route points along the equator.

At this point, readers should consider parallel transport along the sides of a plane equilateral triangle PNQ. It is easy to see that $\boldsymbol{V}(P \to Q)$ is independent of the route taken. Clearly, the difference between the two cases is that the spherical surface is curved while the plane surface is flat. The rule for parallel transport, embodied mathematically in the affine connection coefficients, evidently provides a measure of the curvature of a manifold, and we shall later formulate this precisely. It should be emphasized that a manifold possesses a curvature *only when* it has an affine connection. If it has no connection, then it is neither curved nor flat: the question just does not arise. Finally, returning to figure 2.11, suppose that we had chosen Q to lie close to P and considered only paths contained in a small neighbourhood of the two points. The surface would have been almost indistinguishable from a flat one and the transported vector would have been almost independent of the path. This is consistent with the mathematical expression (2.23). If P has coordinates x^μ and Q is infinitesimally close to P, with coordinates $x^\mu + \mathrm{d}x^\mu$, then we may substitute $\mathrm{d}x^\mu$ for $\delta\lambda \mathrm{d}x^\mu/\mathrm{d}\lambda$, and all reference to the path between P and Q disappears. The affine connection of two-dimensional Euclidean geometry is explored in exercise 2.10.

One of our motivations for introducing the affine connection was to be able to define a meaningful derivative of a vector field. The covariant derivative along a curve was to be defined, using the idea of parallel transport, by (2.22). As we have just seen, it is not actually necessary to specify a curve when P and Q are infinitesimally close. In terms of components, then, let us write $\mathrm{D}V^\mu/\mathrm{d}\lambda = (\mathrm{d}x^\sigma/\mathrm{d}\lambda)\nabla_\sigma V^\mu$ and calculate the covariant derivative $\nabla_\sigma V^\mu$ using (2.22) and (2.23). We find

$$\nabla_\sigma V^\mu = \partial_\sigma V^\mu + \Gamma^\mu{}_{\nu\sigma} V^\nu. \tag{2.24}$$

Notice that the three indices of the connection coefficient have different functions. There are, indeed, important situations in which the connection is *symmetric* in its two lower indices: $\Gamma^{\mu}{}_{\nu\sigma} = \Gamma^{\mu}{}_{\sigma\nu}$. In general, however, it is the last index that corresponds to that of ∇_σ. Since $\mathrm{D}V^\mu/\mathrm{d}\lambda$ and $\mathrm{d}x^\sigma/\mathrm{d}\lambda$ are both vectors, it follows from their transformation laws that the functions $\nabla_\sigma V^\mu$ are the components of a rank $\binom{1}{1}$ tensor, with the transformation law

$$\nabla_{\sigma'} V^{\mu'} = \Lambda^{\sigma}{}_{\sigma'} \Lambda^{\mu'}{}_{\mu} \nabla_\sigma V^\mu. \tag{2.25}$$

From this, we can deduce the transformation law for the connection coefficients themselves, which can be written as

$$\Gamma^{\mu'}{}_{\nu'\sigma'} = \left(\Lambda^{\mu'}{}_{\mu} \Lambda^{\nu}{}_{\nu'} \Lambda^{\sigma}{}_{\sigma'}\right) \Gamma^{\mu}{}_{\nu\sigma} + \Lambda^{\mu'}{}_{\nu} \left(\partial_{\sigma'} \Lambda^{\nu}{}_{\nu'}\right). \tag{2.26}$$

Readers are urged to verify this in detail, bearing in mind that $\partial_{\sigma'}(\Lambda^{\mu'}{}_{\nu}\Lambda^{\nu}{}_{\nu'}) = (\partial_{\sigma'}\Lambda^{\mu'}{}_{\nu})\Lambda^{\nu}{}_{\nu'} + \Lambda^{\mu'}{}_{\nu}(\partial_{\sigma'}\Lambda^{\nu}{}_{\nu'}) = \partial_{\sigma'}(\delta^{\mu'}_{\nu'}) = 0$.

Evidently, the affine connection is not itself a tensor. However, the covariant derivative that contains it acts on any tensor to produce another tensor of one higher covariant rank. So far, we have defined only the covariant derivative of a vector field, which was given in (2.24). The covariant derivative of a scalar field is just the partial derivative, $\nabla_\mu f = \partial_\mu f$, since this is already a vector field. In order for the covariant derivative of a one-form field to be a second-rank tensor field, we must have

$$\nabla_\sigma \omega_\mu = \partial_\sigma \omega_\mu - \Gamma^{\nu}{}_{\mu\sigma} \omega_\nu. \tag{2.27}$$

Notice that the roles of the upper and first lower indices have been reversed, compared with (2.24), and that the sign of the connection term has changed. It is straightforward to check that these changes are vital if this derivative is to transform as a rank $\binom{0}{2}$ tensor field. The covariant derivative of a tensor field of arbitrary rank is

$$\nabla_\sigma T^{\alpha\beta\ldots}{}_{\mu\nu\ldots} = \partial_\sigma T^{\alpha\beta\ldots}{}_{\mu\nu\ldots} + (\text{connection terms}). \tag{2.28}$$

There is one connection term for each index of the original tensor. For each upper index, it is a term like that in (2.24) and for each lower index it is like that in (2.27). Exercise 2.11 invites readers to consider in more detail how these definitions are arrived at.

There is a convenient notation that represents partial derivatives of tensor fields by a comma and covariant derivatives by a semicolon. That is:

$$\partial_\sigma T^{\alpha}{}_{\mu\nu} \equiv T^{\alpha}{}_{\mu\nu,\sigma} \qquad \text{and} \qquad \nabla_\sigma T^{\alpha}{}_{\mu\nu} \equiv T^{\alpha}{}_{\mu\nu;\sigma}. \tag{2.29}$$

2.3.2 Geodesics

As mentioned earlier, a geodesic is, in a sense, a generalization of the straight line of Euclidean geometry. Of course, we can reproduce only those properties of straight lines that make sense in our manifold with its affine connection. For example, the idea that a straight line is the shortest distance between two points will make sense only when we have a metric to measure distances. The idea of a geodesic is that, if we are to walk along a straight line, each step we take must be parallel to the last. Consider, then, the special case of the parallel transport equation (2.23) in which the vector transported from P to Q is the curve's own tangent vector at P: $V^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$. If the curve is a geodesic, the transported vector $\boldsymbol{V}(P \to Q)$ will be proportional to $\boldsymbol{V}(Q)$. Since the vectors have no definite length, the constant of proportionality may well depend on λ, but if P and Q are separated by an infinitesimal parameter distance, it will be only infinitesimally different from 1. So we may write

$$\left.\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\right|_{P\to Q} = [1 - f(\lambda)\delta\lambda]\left.\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\right|_{Q} \tag{2.30}$$

where $f(\lambda)$ is an unknown function. Using this in (2.23) and taking the limit $\delta\lambda \to 0$, we obtain the *geodesic equation*

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} = f(\lambda)\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}. \tag{2.31}$$

A curve $x^\mu(\lambda)$ is a geodesic if and only if it satisfies an equation of this form, where $f(\lambda)$ can be any function.

Remember now that a given path through the manifold can be parametrized in many different ways, each one being regarded as a different curve. It is easy to see that if the curve given by one parametrization is a geodesic, then so is the curve that results from another parametrization of the same path. We need only express the new parameter, say μ, as a function of λ and use the chain rule in (2.31):

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\mu^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\mu}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\mu} = \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^{-2}\left[f(\lambda)\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}^2\mu}{\mathrm{d}\lambda^2}\right]\frac{\mathrm{d}x^\mu}{\mathrm{d}\mu}. \tag{2.32}$$

This has the same form as (2.31) but involves a different function of μ on the right-hand side. In particular, it is always possible to find a parameter for which the right-hand side of (2.32) vanishes. Such a parameter is called an *affine parameter* for the path. It is left as a simple exercise for the reader to show that if λ is an affine parameter, then any parameter that is affinely related to it (that is, it is a linear function $\mu = a\lambda + b$) is also an affine parameter.

2.3.3 The Riemann curvature tensor

We saw in connection with figure 2.11 that parallel transport of a vector between two points along different curves can be used to detect curvature of the manifold. This is because both parallel transport and curvature are properties of the affine

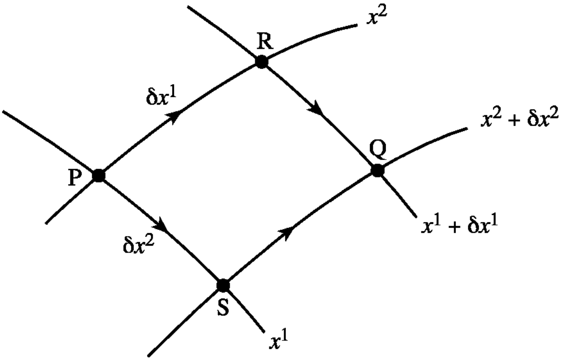


Figure 2.12. Two paths, PRQ and PSQ, for parallelly transporting a vector from P to Q.

connection. The definition of curvature is made precise by the Riemann curvature tensor. Consider two points P and Q with coordinates x^μ and $x^\mu + \delta x^\mu$ respectively, such that $\delta x^\mu = 0$, except for $\mu = 1$ or 2. A region of the (x^1, x^2) surface near these points is shown in figure 2.12. By transporting a vector $\boldsymbol{V}(P)$ to Q via R or S, we obtain at Q the two vectors $\boldsymbol{V}(P \to R \to Q)$ and $\boldsymbol{V}(P \to S \to Q)$. To first order in δx^μ these two vectors are the same, as we have seen. If we expand them to second order, however, they are different, and we obtain an expression of the form

$$V^\mu(P \to S \to Q) - V^\mu(P \to R \to Q) = R^\mu_{\ \nu 12} V^\nu \delta x^1 \delta x^2 + \dots \tag{2.33}$$

where the quantities $R^\mu_{\ \nu 12}$ depend on the connection coefficients and their derivatives. Readers are invited to verify that they are components of the Riemann tensor we are about to define.

It should be clear that the process of transporting the vector from P to Q along the two paths is related to that of taking two derivatives, with respect to x^1 and x^2, in either order. If we act on a vector field with the two covariant derivatives ∇_σ and ∇_τ in succession, the result depends on the order of the two operations; they do not commute. To work out the commutator, we use the definition (2.28), bearing in mind that $\nabla_\sigma V^\mu$ is itself a rank $\binom{1}{1}$ tensor. The result is

$$[\nabla_\sigma, \nabla_\tau] V^\mu \equiv \nabla_\sigma \left(\nabla_\tau V^\mu\right) - \nabla_\tau \left(\nabla_\sigma V^\mu\right) = R^\mu_{\ \nu\sigma\tau} V^\nu + \left(\Gamma^\lambda_{\ \sigma\tau} - \Gamma^\lambda_{\ \tau\sigma}\right) \nabla_\lambda V^\mu \tag{2.34}$$

where

$$R^\mu_{\ \nu\sigma\tau} = \Gamma^\mu_{\ \nu\tau,\sigma} - \Gamma^\mu_{\ \nu\sigma,\tau} + \Gamma^\mu_{\ \lambda\sigma}\Gamma^\lambda_{\ \nu\tau} - \Gamma^\mu_{\ \lambda\tau}\Gamma^\lambda_{\ \nu\sigma}. \tag{2.35}$$

This formidable expression defines the Riemann tensor. As a rank-4 tensor, it has $4^4 = 256$ components! Actually, owing to various symmetry properties, of which the most obvious is antisymmetry in the indices σ and τ, it can be shown that only 80 of these are independent. When $\Gamma^\mu_{\ \nu\sigma}$ is a *metric* connection of the kind to be described in §2.3.5, there is a further symmetry that reduces the number of independent components to 20. Even so, the Riemann tensor is clearly an inconvenient object to deal with. Readers should not panic yet, though. Many of the most important applications of general relativity (including all those to be discussed in this book) do not require the complete Riemann tensor. In practice, we shall need only a simpler tensor derived from it. This is the *Ricci tensor*, defined by contracting two indices of the Riemann tensor:

$$R_{\mu\nu} \equiv R^{\lambda}{}_{\mu\lambda\nu} = \Gamma^{\lambda}{}_{\mu\nu,\lambda} - \Gamma^{\lambda}{}_{\mu\lambda,\nu} + \Gamma^{\lambda}{}_{\sigma\lambda}\Gamma^{\sigma}{}_{\mu\nu} - \Gamma^{\lambda}{}_{\sigma\nu}\Gamma^{\sigma}{}_{\mu\lambda}. \tag{2.36}$$

Although the definition still looks complicated, the components of this tensor can often be calculated with just a little patience, and it is relatively simple to use thereafter.

The second term on the right-hand side of (2.34) involves the antisymmetric part of the affine connection, $\Gamma^{\nu}{}_{\sigma\tau} - \Gamma^{\nu}{}_{\tau\sigma}$, which is called the *torsion* tensor. (Readers should find it instructive to verify, using (2.26) and (2.13) that this really is a tensor, even though $\Gamma^{\nu}{}_{\sigma\tau}$ itself is not.) In most versions of general relativity, it is assumed that spacetime has no torsion. We shall always assume this too, since it makes things much simpler. I do not know, however, of any direct method of testing this experimentally.

Some simple illustrations of the idea of curvature are given in the exercises. These make more obvious sense when we have a metric at our disposal, and we turn to that topic forthwith.

2.3.4 The metric

Yes, we are finally going to give our manifold a metrical structure that will make the notion of length meaningful. To define the infinitesimal distance ds between two points with coordinates x^{μ} and $x^{\mu} + \mathrm{d}x^{\mu}$, we use a generalization of the Pythagoras rule:

$$\mathrm{d}s^2 = g_{\mu\nu}(x)\mathrm{d}x^{\mu}\mathrm{d}x^{\nu}. \tag{2.37}$$

Naturally, we want this distance to be a scalar quantity, independent of our choice of coordinate system, and it is easy to see that the coefficients $g_{\mu\nu}(x)$ must therefore be the components of a rank $\binom{0}{2}$ tensor field. It is called the *metric tensor field* or, for brevity, the 'metric tensor', or simply the 'metric'. Since an antisymmetric part would obviously make no contribution to ds, it is taken to be symmetric in its indices μ and ν. Any finite distance between two points can be uniquely defined only as the length of a specified curve joining them. For the distance between P and Q on a curve $x^{\mu}(\lambda)$, we have the integral

$$s_{PQ} = \int_P^Q \frac{\mathrm{d}s}{\mathrm{d}\lambda}\mathrm{d}\lambda = \int_P^Q \left[g_{\mu\nu}\Big(x(\lambda)\Big)\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\lambda}\frac{\mathrm{d}x^{\nu}}{\mathrm{d}\lambda}\right]^{1/2}\mathrm{d}\lambda. \tag{2.38}$$

In the space of three-dimensional Euclidean geometry, the squared element of distance expressed in Cartesian coordinates is $\mathrm{d}s^2 = (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2$, so the components of the metric tensor in these coordinates are

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.39}$$

The metric tensor has several other geometrical uses, arising from the fact that it serves to define a scalar product of two vectors or vector fields:

$$\boldsymbol{U} \cdot \boldsymbol{V} = U^{\mu}(x) g_{\mu\nu}(x) V^{\nu}(x). \tag{2.40}$$

Clearly, this reduces to the usual ‘dot product’ in Euclidean space. Taking the two vectors to be the same, we get a definition of the magnitude or length of a vector,

$$|\boldsymbol{V}(x)|^2 = g_{\mu\nu}(x) V^{\mu}(x) V^{\nu}(x) \tag{2.41}$$

and we can then define the angle between two vectors by writing

$$g_{\mu\nu} U^{\mu} V^{\nu} = |\boldsymbol{U}||\boldsymbol{V}| \cos\theta. \tag{2.42}$$

A non-Euclidean metric does not necessarily give a positive value for the quantity $|\boldsymbol{V}(x)|^2$, so the lengths and angles defined in this way might turn out to be complex.

When introducing one-forms, I pointed out that the symbol $\boldsymbol{u}\cdot$, which appears in the Euclidean dot product, can be regarded as a linear function that takes a vector as its argument, and is, in fact, a one-form. From the scalar product (2.40), we see that $g_{\mu\nu}$ plays the role of the dot, and that the functions

$$U_{\nu} = U^{\mu} g_{\mu\nu} \tag{2.43}$$

are the components of a unique one-form corresponding to the vector $\boldsymbol{U}$. The metric tensor is said to *lower the index* of the vector to produce a one-form. In the same way, the metric associates a unique vector with each one-form ω: it is the vector whose corresponding one-form is ω. Actually, this assumes that the metric is non-singular. That is, it has an inverse matrix $g^{\mu\nu}$, whose elements are the components of a rank $\binom{2}{0}$ tensor field, such that

$$g_{\mu\sigma} g^{\sigma\nu} = \delta^{\nu}_{\mu}. \tag{2.44}$$

The geometrical properties of the metric would be rather peculiar if this were not so, and the existence of the inverse is sometimes included as part of the definition of a metric. So long as the inverse metric does exist, we can say that it *raises the index* of a one-form to produce a vector:

$$\omega^{\mu} = g^{\mu\nu} \omega_{\nu}. \tag{2.45}$$

In fact, any index of any tensor can be raised or lowered in this way. Since $g_{\mu\nu}$ is symmetric, it does not matter which of its indices is contracted.

Now that we have a metric tensor at our disposal, it is clearly possible in practice to regard vectors and one-forms as different versions of the same thing—hence the terms contravariant and covariant vector. In Euclidean geometry, we do not notice the difference, as long as we use Cartesian coordinates, because the metric tensor is just the unit matrix. In non-Cartesian coordinates, the metric tensor is not the unit matrix, and some consequences of this are explored in the exercises. Does this mean that there is, after all, no real distinction between vectors and one-forms, or between the contravariant and covariant versions of other tensors? This depends on our attitude towards the metric. In the relativistic theory of gravity, the metric embodies information about gravitational fields, and different metrics may represent different, but equally possible, physical situations. The relation between the contravariant and covariant versions of a given physical quantity depends on the metric, and it is legitimate to ask which version is intrinsic to the quantity itself and which is a compound of information about the quantity itself and about the metric. To decide this, we must ask what kind of tensor would be used to represent the quantity in question were a metric not available. For example, the Riemann tensor that appears in (2.34) has an index μ, which is in the upper position because it originates from parallel transport of a vector, and two indices σ and τ that must be in the lower position because they label directions along which the vector is being differentiated. Since metrical notions are taken for granted in much of our physical thinking, though, the answer to this may not always be obvious. If, as in Euclidean geometry, the metric is taken to be fixed and unalterable, then such questions need not arise.

2.3.5 The metric connection

Now that the magnitude of a vector and the angle between two vectors have acquired definite meanings, it is natural to demand that the rule for parallel transport should be consistent with them. Thus, if two vectors are transported along a curve, each one remaining parallel to itself, then the angle between them should remain constant. This requirement leads to a relation between the metric and the affine connection that we shall now derive. Consider a curve $x^\mu(\lambda)$ passing through the point P and two vectors $\boldsymbol{V}$ and $\boldsymbol{W}$ at P. We can define a vector field $\boldsymbol{V}(x)$ such that its value at any point Q on the curve is equal to the transported vector $\boldsymbol{V}(P \to Q)$, and a similar vector field $\boldsymbol{W}(x)$. If $\boldsymbol{U}$ is the tangent vector to the curve, then $U^\sigma \nabla_\sigma V^\mu$ is the covariant derivative of V^μ along the curve. It is given by the expression (2.22) and is clearly equal to zero, as is the corresponding derivative of $\boldsymbol{W}$. The consistency condition we want to impose is that the scalar product $g_{\mu\nu} V^\mu W^\nu$ has the same value everywhere along the curve. Recalling that the covariant derivative of a scalar field is equal to the ordinary derivative, we may express this condition as

$$U^\sigma \nabla_\sigma (g_{\mu\nu} V^\mu W^\nu) = 0. \tag{2.46}$$

Now, the covariant derivative of a product of tensors obeys the same Leibniz (or product) rule as an ordinary derivative:

$$\nabla_\sigma(g_{\mu\nu}V^\mu W^\nu) = (\nabla_\sigma g_{\mu\nu})V^\mu W^\nu + g_{\mu\nu}(\nabla_\sigma V^\mu)W^\nu + g_{\mu\nu}V^\mu(\nabla_\sigma W^\nu). \quad (2.47)$$

Readers may verify this explicitly or turn to exercise 2.11 for some further enlightenment. If we use this in (2.46), the last two terms vanish and our condition becomes $U^\sigma(\nabla_\sigma g_{\mu\nu})V^\mu W^\nu = 0$. This must hold for any three vectors $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{W}$, and therefore the covariant derivative of $g_{\mu\nu}$ must be zero:

$$\nabla_\sigma g_{\mu\nu} = g_{\mu\nu,\sigma} - \Gamma^\tau{}_{\mu\sigma}g_{\tau\nu} - \Gamma^\tau{}_{\nu\sigma}g_{\mu\tau} = 0. \quad (2.48)$$

This is sometimes expressed by saying that the metric is 'covariantly constant'. By combining this equation with two others obtained by renaming the indices, we get

$$g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma} = (\Gamma^\tau{}_{\sigma\nu} - \Gamma^\tau{}_{\nu\sigma})g_{\tau\mu} + (\Gamma^\tau{}_{\sigma\mu} - \Gamma^\tau{}_{\mu\sigma})g_{\tau\nu} + (\Gamma^\tau{}_{\mu\nu} + \Gamma^\tau{}_{\nu\mu})g_{\tau\sigma}. \quad (2.49)$$

Assuming, as we discussed above, that the connection is symmetric in its lower indices, the first two terms on the right-hand side vanish. Then, on multiplying by $g^{\lambda\sigma}$, we find that this symmetric connection is completely determined by the metric:

$$\Gamma^\lambda{}_{\mu\nu} = \tfrac{1}{2}g^{\lambda\sigma}(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma}). \quad (2.50)$$

When $\Gamma^\lambda{}_{\mu\nu}$ is used to denote this expression, it is often called a *Christoffel symbol*. This metric connection expresses the definition of parallelism that is implied by the metric. In principle, there is no reason why a manifold should not possess one or more affine connections that would be quite independent of the metric. Indeed, it might also possess several different metrics. In such a case, there would exist several different kinds of 'distance' and several different meanings of 'parallel'. It appears, however, that a single metric and its associated connection given by (2.50) are sufficient to describe the properties of space and time as we know them.

Finally, we can now construct a scalar quantity that gives a measure of curvature (though it obviously contains much less information than the full Riemann tensor). The *Ricci curvature scalar* R is defined by

$$R = g^{\mu\nu}R_{\mu\nu} \quad (2.51)$$

and its interpretation in terms of a 'radius of curvature' is explored in exercise 2.15.

2.4 What is the Structure of Our Spacetime?

We have now invested considerable effort in understanding the mathematical nature of the affine and metrical structures that give precise meaning to our

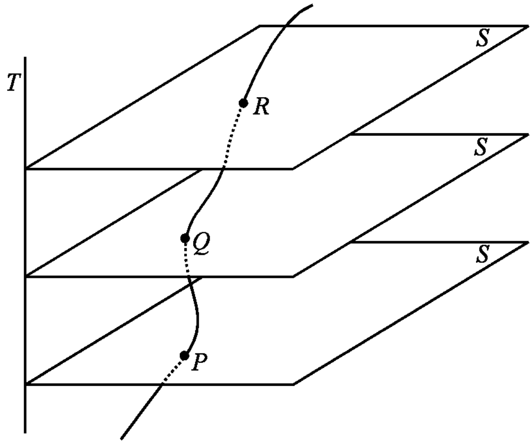


Figure 2.13. Fibre bundle structure of Galilean spacetime and the trajectory of a particle moving through it. Each fibre is a copy of three-dimensional Euclidean space S, which possesses a metric for measuring distances. The base manifold T has its own metric for measuring time intervals. There is no unique way of measuring the 'length' of the particle's trajectory.

intuitive geometrical ideas. The question naturally arises, what are the particular structures that occur in our real, physical space and time? Let us first consider what kind of an answer is needed.

Before Einstein's theories of relativity, it had seemed obvious that the geometry of space was that described by Euclid. (The logical possibility of non-Euclidean geometry had, however, been investigated rather earlier by Gauss, Bolyai, Lobachevski, Riemann and others. The history of this subject is nicely summarized by Weinberg (1972).) The Galilean spacetime that incorporates Euclidean geometry does not have exactly the kind of metrical structure we have been considering. It is a combination (in mathematical jargon, a *direct product*) of two manifolds T (time) and S (space), each of which has its own metric. This structure, illustrated in figure 2.13, is called a *fibre bundle*. It has a base manifold, T, to each point of which is attached a fibre. Each fibre is a copy of the three-dimensional Euclidean space S. A curve such as PQR passing through the spacetime has no well-defined length, although its projection onto one of the fibres does have a definite length l and its projection onto T spans a definite time interval t.

The big difference between Galilean spacetime and the spacetimes of Einstein's theories is that the latter are *metric spaces* (or, more accurately, *manifolds-with-metrics*). That is, the spacetime is a manifold in which a single metric tensor field defines, as we saw in our initial survey, the arc length of any curve. This 'length' is a combination of temporal and spatial intervals, but there is no unique way in which the two can be separated. There is, of course, a profound difference between space and time as we experience them, and we shall discuss in later chapters how this difference fits in with the mathematics.

An important similarity between Galilean spacetime and the Minkowski spacetime of special relativity is that their metrical properties are assumed to be known *a priori*, as specified either by (2.39) or by (2.8). Readers may be puzzled to see that the spatial components in (2.8) have changed sign relative to (2.39). This is purely a matter of convention: the squared proper time intervals in (2.3) or (2.6) are taken to be positive if the separation of two events in time is greater than $1/c$ times their spatial separation, and negative otherwise. (Since proper time intervals are scalar quantities, having the same values in all frames of reference, this distinction is also independent of the frame in which the time and distance measurements are made.) If we chose to think in terms of proper distance rather than proper time, the opposite convention would be more natural, and every component in (2.8) would have the opposite sign. In fact, both conventions are used in the literature, although the one we are using is somewhat more popular amongst high-energy physicists than amongst relativity theorists.

The crux of the general-relativistic theory of gravity is that neither of these simple assumptions about the metric tensor is in fact correct. Indeed, the most important conceptual step we have taken in this chapter is to recognize that the metric tensor is not an intrinsic part of the spacetime manifold, but rather an object that lives in the manifold. It is the same sort of thing as an electric or magnetic field. Electric and magnetic fields vary with position and time in accordance with definite physical laws, which relate them to distributions of charged particles and currents. In the same way, the metric tensor field can be expected to vary in accordance with its own laws of motion and to depend on the distribution of matter. So far, we have no idea what the laws of motion for the metric tensor field are. Electromagnetic fields are easy to produce and control under laboratory conditions, and the laws that govern them were, for the most part, inferred from comprehensive experimental investigations. In contrast, the gravitational forces that are the observable manifestations of the metric tensor field are extremely weak, unless they are produced by bodies of planetary size, and there is little hope of deducing the laws that govern them from a series of controlled experiments. What Einstein did was to guess at what these laws might be, assuming that they would be reasonably similar to other known laws of physics. After one or two false guesses, he arrived at a set of equations, the *field equations* of general relativity, which are consistent with the most precise astronomical observations that it has so far been possible to make.

With the benefit of hindsight, it is possible to see that these equations and all the other laws of classical (non-quantum-mechanical) physics can be deduced in exactly the same way from a single basic principle, called an *action principle*. This seems to me to be most satisfactory. I should be vastly more satisfied if I could explain why an action principle rather than something else is what actually works, but I cannot imagine how that would be done. (It *is* possible to derive the classical action principle from what amounts to a quantum-mechanical version of the same thing, but that is only to rephrase the question!) At this point, then, I propose to interrupt our study of geometry to examine how classical physics works in Galilean and Minkowski spacetimes. This is an important topic in its own right, because classical physics and the simple spacetimes often provide excellent approximations to the real world. In the course of understanding them, however, we shall also meet the action principle, whereupon we shall be equipped to embark upon general relativity and the theory of gravity.

Exercises

2.1. Consider two coordinate systems S and S' whose spatial Cartesian axes lie in the same three directions. The origin of S' moves with constant velocity $\boldsymbol{v}$ relative to S, and the origins of S and S' coincide at $t = t' = 0$. Assume that the relation between the two sets of coordinates is linear and that space is isotropic. The most general form of the transformation law can then be written as

$$\boldsymbol{x}' = \alpha\left[(1 - \lambda v^2)\boldsymbol{x} + (\lambda \boldsymbol{v}\cdot\boldsymbol{x} - \beta t)\boldsymbol{v}\right] \qquad t' = \gamma\left[t - (\delta/c^2)\boldsymbol{v}\cdot\boldsymbol{x}\right]$$

where α, β, γ, δ and λ are functions of v^2. For the case that $\boldsymbol{v}$ is in the positive x direction, write out the transformations for the four coordinates. Write down the trajectory of the S' origin as seen in S and that of the S origin as seen in S' and show that $\beta = 1$ and $\alpha = \gamma$. Write down the trajectories seen in S and S' of a light ray emitted from the origin at $t = t' = 0$ that travels in the positive x direction, assuming that it is observed to travel with speed c in each case. Show that $\delta = 1$. The transformation from S' to S should be the same as the transformation from S to S', except for the replacement of v by $-v$. Use this to find γ. By considering the equation of the spherical wavefront of a light wave emitted from the origin at $t = t' = 0$, complete the derivation of the Lorentz transformation (2.2).

2.2. Two coordinate frames are related by the Lorentz transformation (2.2). A particle moving in the x direction passes their common origin at $t = t' = 0$ with velocity u and acceleration a as measured in S. Show that its velocity and acceleration as measured in S' are

$$u' = \frac{u - v}{1 - uv/c^2} \qquad a' = \frac{(1 - v^2/c^2)^{3/2}}{(1 - uv/c^2)^3}\,a.$$

2.3. A rigid rod of length L is at rest in S', with one end at $x' = 0$ and the other at $x' = L$. Find the trajectories of the two ends of the rod as seen in S and show that the length of the rod as measured in S is L/γ, where $\gamma = (1 - v^2/c^2)^{-1/2}$. This is the *Fitzgerald contraction*. If the rod lies along the y' axis of S', what is its apparent length in S? A clock is at rest at the origin of S'. It ticks at $t' = 0$ and again at $t' = \tau$. Show that the interval between these ticks as measured in S is $\gamma\tau$. This is *time dilation*.

2.4. As seen in S, a signal is emitted from the origin at $t = 0$, travels along the x axis with speed u, and is received at time τ at $x = u\tau$. Show that, if $u > c^2/v$ then, as seen in S', the signal is received before being sent. Show that if such paradoxes are to be avoided, no signal can travel faster than light.

2.5. A wheel has a perfectly rigid circular rim connected by unbreakable joints to perfectly rigid spokes. When measured at rest, its radius is r and its circumference is $2\pi r$. When the wheel is set spinning with angular speed ω, what, according to exercise 2.3, is the apparent circumference of its rim and the apparent length of its spokes? What is the speed of sound in a solid material of density ρ whose Young's modulus is Y? Is the notion of a perfectly rigid material consistent with the conclusion of exercise 2.4?

2.6. Consider the following three curves in the Euclidean plane with Cartesian coordinates x and y: (i) $x = 2\sin\lambda$, $y = 2\cos\lambda$, $0 \leq \lambda < 2\pi$; (ii) $x = 2\cos(s/2)$, $y = 2\sin(s/2)$, $0 \leq s < 4\pi$; (iii) $x = 2\cos(\mathrm{e}^{\mu})$, $y = 2\sin(\mathrm{e}^{\mu})$, $-\infty < \mu \leq \ln(2\pi)$. Show that all three curves correspond to the same path, namely a circle of radius 2. Show that λ and s are affinely related. What is the special significance of s? Find the components of the tangent vectors to each curve. Compare the magnitudes and directions of the three tangent vectors at various points on the circle. What is special about the tangent vectors to curve (ii)?

2.7. Consider a four-dimensional manifold and a specific system of coordinates x^{μ}. You are given four functions, $a(x^{\mu})$, $b(x^{\mu})$, $c(x^{\mu})$ and $d(x^{\mu})$. Can you tell whether these are (i) four scalar fields, (ii) the components of a vector field, (iii) the components of a one-form field or (iv) none of these? If not, what further information would enable you to do so?

2.8. In the Euclidean plane, with Cartesian coordinates x and y, consider the vector field $\boldsymbol{V}$ whose components are $V^x = 2x$ and $V^y = y$, and the one-form field ω_f which is the gradient of the function $f = x^2 + y^2/2$. Show that in any system of Cartesian coordinates $x' = x\cos\alpha + y\sin\alpha$, $y' = y\cos\alpha - x\sin\alpha$, where α is a fixed angle, the components of ω_f are identical to those of $\boldsymbol{V}$. In polar coordinates (r, θ), such that $x = r\cos\theta$ and $y = r\sin\theta$, show that $\boldsymbol{V}$ has components $(r(1 + \cos^2\theta), -\sin\theta\cos\theta)$ while ω_f has components $(r(1 + \cos^2\theta), -r^2\sin\theta\cos\theta)$. Note that the 'gradient vector' defined in elementary vector calculus to have the components $(\partial f/\partial r, r^{-1}\partial f/\partial\theta)$ does not correspond to either $\boldsymbol{V}$ or ω_f.

2.9. Given a rank $\binom{a}{b}$ tensor, show that the result of contracting any upper index with any lower index is a rank $\binom{a-1}{b-1}$ tensor.

2.10. In the Euclidean plane, parallel transport is defined in the obvious way. If, in

Cartesian coordinates, the components of $\boldsymbol{V}(P)$ are (u, v), then the components of $\boldsymbol{V}(P \to Q)$ are also (u, v). Thus, the affine connection coefficients in Cartesian coordinates are all zero. Work out the matrices $\Lambda^{\mu'}{}_{\mu}$ for transforming between Cartesian and polar coordinates related by $x = r\cos\theta$ and $y = r\sin\theta$. Show that in polar coordinates, the only non-zero connection coefficients are $\Gamma^{r}{}_{\theta\theta} = -r$ and $\Gamma^{\theta}{}_{r\theta} = \Gamma^{\theta}{}_{\theta r} = 1/r$. Let P and Q be the points with Cartesian coordinates $(a, 0)$ and $(a\cos\alpha, a\sin\alpha)$ respectively, and let $\boldsymbol{V}(P)$ have Cartesian components $(1, 0)$. Using polar coordinates and parallel transport around the circle of radius a centred at the origin and parametrized by the polar angle θ, show that $\boldsymbol{V}(P \to Q)$ has polar components $(\cos\alpha, -a^{-1}\sin\alpha)$. By transforming this result, verify that $\boldsymbol{V}(P \to Q)$ has Cartesian components $(1, 0)$. [N.B. The notation here is intended to be friendly: if, say, $x^1 = r$ and $x^2 = \theta$, then $\Gamma^{r}{}_{\theta\theta}$ means $\Gamma^{1}{}_{22}$ and so on.]

2.11. The covariant derivatives of tensors of arbitrary rank can be defined recursively by the following rules: (i) for a scalar field f, we take $\nabla_\sigma f = \partial_\sigma f$; (ii) the covariant derivative of a vector field is given by (2.24); (iii) the covariant derivative of a rank $\binom{a}{b}$ tensor is a tensor of rank $\binom{a}{b+1}$; (iv) for any two tensors A and B, the Leibniz rule $\nabla_\sigma(AB) = (\nabla_\sigma A)B + A(\nabla_\sigma B)$ holds. By considering the fact that $\omega(\boldsymbol{V}) = \omega_\mu V^\mu$ is a scalar field, show that the covariant derivative of a one-form is given by (2.27). Convince yourself that the recursive definition leads to (2.28) for an arbitrary tensor field.

2.12. In the Euclidean plane, consider the straight line $x = a$. Using $\lambda = y$ as a parameter, show, in both Cartesian and polar coordinates, that the geodesic equation (2.31) is satisfied and that λ is an affine parameter. Repeat the exercise using both affine and non-affine parameters of your own invention.

2.13. Write down the components of the metric tensor field of the Euclidean plane in the polar coordinates of exercise 2.8. Show, using both Cartesian and polar coordinates, that the vector $\boldsymbol{V}$ is obtained by raising the indices of ω_f and *vice versa*. Show that $|\boldsymbol{V}|^2 = \omega_f(\boldsymbol{V})$. What is the magnitude of the 'gradient vector'? How does it involve the metric? Can a 'gradient vector' be defined in a manifold with a non-Euclidean metric, or in a manifold that possesses no metric?

2.14. Show that the affine connection of exercise 2.10 is the metric connection.

2.15. In three-dimensional Euclidean space, define polar coordinates in the usual way by $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$ and $z = r\cos\theta$. The spherical surface $r = a$ is called a 2-sphere, and the angles θ and ϕ can be used as coordinates for this two-dimensional curved surface. Show that the line element on the sphere is $ds^2 = a^2(d\theta^2 + \sin^2\theta\, d\phi^2)$. Show that the only non-zero coefficients of the metric connection are $\Gamma^{\theta}{}_{\phi\phi} = -\sin\theta\cos\theta$ and $\Gamma^{\phi}{}_{\theta\phi} = \Gamma^{\phi}{}_{\phi\theta} = \cot\theta$. Show that the Ricci tensor is diagonal, with elements $R_{\theta\theta} = 1$ and $R_{\phi\phi} = \sin^2\theta$, and that the Ricci scalar is $R = 2/a^2$.

Chapter 3

Classical Physics in Galilean and Minkowski Spacetimes

This chapter is mostly about classical mechanics. By ‘classical’, I mean to indicate that we are not yet going to take any account of quantum mechanics. (In the literature, ‘classical’ is sometimes used to mean that no account is taken of special relativity either, and sometimes also to describe any venerable theory that has been superseded by a more ‘modern’ one.) I shall actually be assuming that readers already have a fair understanding of the elementary aspects of Newtonian mechanics: for example, we shall not spend time developing techniques for calculating the trajectories of projectiles or planetary orbits, important though these topics undoubtedly are. The aim of this chapter is to set out the mathematics of classical mechanics in a way that makes clear the nature of the basic physical laws embodied in it and which, to a large extent, will enable us to see the principles of general relativity and of the quantum theory as natural generalizations of these laws. In a later chapter, this mathematical description will also help us towards setting up a statistical description of the macroscopic behaviour of large assemblages of particles.

There is, of course, nothing final or unalterable about the ‘laws’ of physics as they appear to physicists at any particular time. It is possible, however, to identify two mathematical ideas which lie at the heart of all theories that have so far had success in describing how the world is at a fundamental level. The first is a function called the *action* which, as we shall soon see, summarizes all the equations of motion for a given system. It is easy to invent equations of motion that cannot be summarized in this way. For example, equations that involve dissipative effects such as friction usually cannot be. These effects, however, can be understood as arising only on a macroscopic scale, and the fundamental equations that apply at the microscopic level do seem to be derivable from an action. Why this should be so, I do not know.

The action is fundamental to both classical and quantum theories, although in somewhat different guises. It is a function of all the dynamical variables (in the classical mechanics of particles, the positions and velocities of all the particles) that are needed to specify the state of a system. Once we know what this function is, we know what the laws of motion are, but the only way of finding this out is by guesswork. It seems that the possibilities amongst which we have to choose are quite considerably restricted by a variety of *symmetries* that are respected by nature. This is the second of the ideas mentioned above, and the role of symmetry in theoretical physics will be a recurring theme. Symmetry is, of course, an aesthetically pleasing feature of any theory, and this has come to weigh heavily with many physicists. At the same time, it is not really clear why nature should share our aesthetic tastes, if indeed she really does. But symmetry is more than a theoretician's fancy. As we shall soon discover, every symmetry leads to a conservation law, the best-known examples being, perhaps, the conservation of energy, momentum and electric charge. These conservation laws are amenable to quite rigorous experimental checks and, conversely, the empirical discovery of conserved quantities may point to new symmetries that should be incorporated in our mathematical models.

These, then, are the issues to which the present chapter is primarily addressed.

3.1 The Action Principle in Galilean Spacetime

The basic problem we set ourselves in classical mechanics is, given the state of a system at some initial time, to predict what its state will be at some later time. If we can do this correctly or if, at least, we are satisfied that only computational difficulties stand in the way of our doing it, then we feel that we understand how the system works. We shall be concerned more or less exclusively with systems consisting of particles that are small enough to be considered as points. Large rigid bodies can be treated as being composed of such particles and introduce no new questions of principle.

Let us consider first what information we need to specify uniquely the instantaneous state of such a system. It is normally taken for granted that we have to know the positions and velocities of all the particles—whether these are given in Cartesian coordinates for each particle, in polar coordinates, in terms of relative positions and velocities for some of the particles, etc. does not matter. But why is this? A snapshot of the system can be completely described by giving just the positions of the particles. Evidently, this is not enough, but if we go on to specify the velocities, then why not the accelerations and higher-order time derivatives as well? By saying that the state of the system is uniquely specified, we imply that, given the equations of motion, any future state is uniquely determined. The equations of motion come simply from Newton's second law, which gives a set of second-order differential equations for the positions of the particles as functions of time. They have unique solutions if the initial positions and velocities are given. I emphasize this point because I am going to illustrate the role of symmetries by using them to *derive* Newton's second law, and I want to be clear about the assumptions that are needed to do this. The first assumption is that the state of the system is uniquely specified by giving the positions and velocities, and it is more or less equivalent to assuming that the equations of motion will be of second order in the time derivatives. I do not know of any justification for this beyond the fact that it works.

At this point, we must introduce the action principle. As a simple example, consider a single particle in a Galilean spacetime with one spatial dimension. If its mass is m and it has potential energy $V(x)$, then Newton's law gives

$$m\ddot{x} = -\mathrm{d}V/\mathrm{d}x. \tag{3.1}$$

This is equivalent to the statement that the quantity

$$S = \int_{t_1}^{t_2} \left[\tfrac{1}{2}m\dot{x}^2 - V(x)\right]\mathrm{d}t \tag{3.2}$$

called the *action*, is stationary with respect to variations in the path $x(t)$. That is to say, if $x(t)$ is the actual path of the particle, and we imagine changing it by a small but otherwise arbitrary amount, $x(t) \rightarrow x(t) + \delta x(t)$, then the resulting first-order change in S is zero:

$$\delta S = \int_{t_1}^{t_2} [m\dot{x}\delta\dot{x} - (\mathrm{d}V/\mathrm{d}x)\delta x]\,\mathrm{d}t = 0. \tag{3.3}$$

To be precise, we must choose $\delta x(t)$ to vanish at t_1 and t_2. Then, taking into account that $\delta\dot{x} = \mathrm{d}(\delta x)/\mathrm{d}t$, we may integrate the first term by parts, giving

$$\int_{t_1}^{t_2} [m\ddot{x} + (\mathrm{d}V/\mathrm{d}x)]\,\delta x\,\mathrm{d}t = 0. \tag{3.4}$$

Since $\delta x(t)$ is an arbitrary function, the expression in square brackets must be zero, and in this way we recover the equation of motion (3.1). The integrand in (3.2) is called the *Lagrangian* and in this case it can be identified as (kinetic energy − potential energy).

In general, for a system of N particles in three-dimensional space, its instantaneous state is specified by a set of $3N$ quantities $\{q_i\}$, called *generalized coordinates*, which may be distances, angles, or any other quantities that serve to specify all the positions, together with the $3N$ *generalized velocities* $\{\dot{q}_i\}$. Then the Lagrangian may be a function of all $6N$ of these quantities and of time, $L = L(\{q_i\}, \{\dot{q}_i\}, t)$. By repeating the above calculation, but allowing for independent variations in all the coordinates, readers may easily verify that the resulting equations of motion are the $3N$ equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}. \tag{3.5}$$

These are called the *Euler–Lagrange* equations. The quantity $p_i = \partial L/\partial\dot{q}_i$ is called the *generalized momentum* conjugate to the coordinate q_i, and $\partial L/\partial q_i$ is the *generalized force*. The rate of change of a generalized momentum is thus a generalized force and, by choosing the Lagrangian function correctly, these equations can be made to reproduce those given by Newton's law.

Suppose, however, we do not assume Newton's law to be valid. Can we discover what the Lagrangian is on *a priori* grounds? In fact, quite a lot can be discovered by considering *spacetime symmetries*, as we shall now see. Consider first the case of a single, isolated particle. Since it is free from external influences, its equation of motion can depend only on the structure of spacetime itself: any symmetry of this structure must also be a symmetry of the equation of motion. In Galilean spacetime, there are three quite obvious symmetries, which place definite constraints on the Lagrangian.

(i) *Invariance under time translations.* In terms of the geometrical ideas in the last chapter, Galilean time has its own metric, which gives a definite quantitative meaning to time intervals. We assume that the time coordinate t, as well as labelling instants of time, is a linear measure of time. This means that, given any other parameter t' that labels instants of time (say, the readings of an imperfect clock), there is a temporal metric tensor with a single component $g(t')$ such that $\mathrm{d}t^2 = g(t')\mathrm{d}t'^2$. In terms of t itself, $g = 1$, so there is nothing to distinguish one moment from any other. Thus, the equation of motion of an isolated particle must be the same at any instant, and therefore L cannot depend explicitly on time. Another way of saying this is that L is invariant under the coordinate transformation that shifts or 'translates' the origin of time measurement by an amount t_0: $L(\boldsymbol{x}, \dot{\boldsymbol{x}}, t + t_0) = L(\boldsymbol{x}, \dot{\boldsymbol{x}}, t)$ so L is independent of t, which can be omitted.

(ii) *Invariance under spatial translations.* In Cartesian coordinates, the Pythagoras rule for finding the length of a segment of a curve is unchanged by a translation of the origin $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{x}_0$ or, in the terminology of the last chapter, the spatial metric tensor (2.39) is unchanged. By the same reasoning as above, we conclude that $L(\boldsymbol{x} + \boldsymbol{x}_0, \dot{\boldsymbol{x}}) = L(\boldsymbol{x}, \dot{\boldsymbol{x}})$ or that L must be a function of $\dot{\boldsymbol{x}}$ only.

(iii) *Invariance under rotations.* Similarly, the Pythagoras rule or the metric tensor is unchanged by a rotation to a new Cartesian coordinate system. Therefore, L must be invariant under rotations. This means that it cannot depend on individual components of $\dot{\boldsymbol{x}}$ but only on the magnitude $|\dot{\boldsymbol{x}}| = (\dot{\boldsymbol{x}} \cdot \dot{\boldsymbol{x}})^{1/2}$ which is unchanged by rotations.

In order to tie down the Lagrangian completely, we have to assume a further symmetry:

(iv) *Invariance under Galilean transformations.* This is the assumption that the equation of motion has the same form in two frames of reference that have a constant relative velocity $\boldsymbol{v}$. The interpretation of this symmetry in terms of the geometry of Galilean spacetime is somewhat obscure, although it can be understood as a limiting case of the invariance under Lorentz transformations that applies in Minkowski spacetime. Clearly, it involves assuming the existence of a privileged class of unaccelerated or inertial frames of reference in which the equation of motion has a special form. We found above that L can depend only on $|\dot{\boldsymbol{x}}|$, and it is now convenient to express L as a function of the variable $X = \frac{1}{2}|\dot{\boldsymbol{x}}|^2$. If we choose the generalized coordinates in (3.5) to be Cartesian, then the equation of motion can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{\boldsymbol{x}}\frac{\mathrm{d}L}{\mathrm{d}X}\right) = \ddot{\boldsymbol{x}}\frac{\mathrm{d}L}{\mathrm{d}X} + \dot{\boldsymbol{x}}(\ddot{\boldsymbol{x}}\cdot\dot{\boldsymbol{x}})\frac{\mathrm{d}^2L}{\mathrm{d}X^2} = 0. \tag{3.6}$$

If we make a Galilean transformation, replacing $\boldsymbol{x}$ by $\boldsymbol{x} - \boldsymbol{v}t$, then $\dot{\boldsymbol{x}}$ and X are changed, but $\ddot{\boldsymbol{x}}$ is not. To ensure that the form of (3.6) remains unchanged, we must take L to be such that $\mathrm{d}L/\mathrm{d}X$ is simply a constant, which means that $\mathrm{d}^2L/\mathrm{d}X^2 = 0$. The constant is, of course, what we usually call the mass of the particle, and the Lagrangian has turned out to be just the kinetic energy, $L = \frac{1}{2}m\dot{\boldsymbol{x}}^2$, as it ought to be.

The Lagrangian for a system of non-interacting particles will clearly be the sum of the kinetic energies of all the particles. If the particles interact with each other, it will contain further terms to account for the forces. To maintain the invariance under space translations and Galilean transformations, we can include in these additional terms only functions of the separations $\boldsymbol{r}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$ and relative velocities $\dot{\boldsymbol{r}}_{ij} = \dot{\boldsymbol{x}}_i - \dot{\boldsymbol{x}}_j$ of pairs of particles, so the general form of the Lagrangian is

$$L = \sum_i \tfrac{1}{2}m_i\dot{\boldsymbol{x}}_i^2 - V(\{\boldsymbol{r}_{ij}\}, \{\dot{\boldsymbol{r}}_{kl}\}). \tag{3.7}$$

To maintain rotational invariance, V can depend only on scalar quantities constructed from these vectors, $\boldsymbol{r}_{ij} \cdot \boldsymbol{r}_{kl}$, $(\boldsymbol{r}_{ij} \times \boldsymbol{r}_{kl}) \cdot \dot{\boldsymbol{r}}_{mn}$ and so on, but no more can be said *a priori* about the function V, unless we can identify other symmetries that apply to specific systems.

Our original example (3.2) is not of this form and, unless V is a trivial constant, $V(x + x_0)$ does not equal $V(x)$. If our symmetry arguments are correct, then a Lagrangian of this kind can arise only when the potential is produced by some external system whose own behaviour is not taken properly into account. This may well be an excellent approximation. For example, the motion of a small object (mass m, position $\boldsymbol{x}$) near the Earth (mass M, position $\boldsymbol{X}$) would, according to Newtonian gravity, be described by a Lagrangian of the form (3.7), with $V = -GmM/|\boldsymbol{x} - \boldsymbol{X}|$. For many purposes, we can simply take the Earth to be fixed, say at $\boldsymbol{X} = \boldsymbol{0}$, so that V becomes a function of $\boldsymbol{x}$ only. For the small object on its own, translational invariance does not hold because of the presence of the Earth, but for the combined system of object + Earth, translational invariance does hold, so long as we neglect any influence of the rest of the universe. Thus, we expect the symmetries to be valid for any *isolated* system.

3.2 Symmetries and Conservation Laws

We saw above that the symmetry of invariance under time translations implied that the Lagrangian could not depend explicitly on time. Therefore, all the time dependence of L is through the generalized coordinates and velocities, and we may write

$$\frac{\mathrm{d}L}{\mathrm{d}t}=\sum_i\left(\frac{\mathrm{d}q_i}{\mathrm{d}t}\frac{\partial L}{\partial q_i}+\frac{\mathrm{d}\dot{q}_i}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right)=\sum_i\left(\dot{q}_i\frac{\partial L}{\partial q_i}+\ddot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right). \tag{3.8}$$

When the functions $q_i(t)$ represent the actual trajectories of the particles, and therefore obey the equations of motion (3.5), this becomes

$$\frac{\mathrm{d}L}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_i \dot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right) \tag{3.9}$$

which shows that $\mathrm{d}E/\mathrm{d}t=0$, where

$$E=\sum_i \dot{q}_i\frac{\partial L}{\partial \dot{q}_i}-L. \tag{3.10}$$

This quantity, therefore, is conserved: it is a 'constant of the motion'. When the Lagrangian is that in (3.2), we see that E is the total energy. In general, since the concept of energy is useful only because of the conservation law, we might as well regard (3.10) as *defining* the energy of the system. (There are awkward cases in which other definitions of energy give a different result from (3.10), but we shall not be meeting them.) Thus, *if the Lagrangian does not depend explicitly on time, or is invariant under time translations, then energy is conserved.* As discussed above, we would expect this symmetry, and thus the conservation law, to hold for any isolated system. This seems to me to be a remarkable and most satisfying result. Far from depending on the details of forces that act within any particular system, the law of conservation of energy is simply a consequence of the fact that one instant of time is as good as any other, as far as the laws of physics are concerned. This 'fact' might, indeed, have seemed to be more or less self evident, had we not encountered in the last chapter the idea that spacetime geometry, as embodied in the metric, might after all vary from one time and place to another. In Galilean or Minkowski spacetime, this does not happen, but we might anticipate that conservation of energy will not be so straightforward an idea in the context of general relativity.

A variety of other conservation laws can be deduced from symmetry or invariance properties of the Lagrangian. Mathematically, this works in the following way. We replace the coordinates q_i by $q_i+\epsilon f_i$ and the velocities by $\dot{q}_i+\epsilon\mathrm{d}f_i/\mathrm{d}t$, where each f_i is a function of the coordinates, velocities and time, and ϵ is a small, constant parameter. The Lagrangian can be expanded as a Taylor series in ϵ:

$$L\left(q_i + \epsilon f_i, \dot{q}_i + \epsilon \frac{\mathrm{d}f_i}{\mathrm{d}t}, t\right) = L(q_i, \dot{q}_i, t) + \epsilon \sum_j \left(\frac{\partial L}{\partial q_j} f_j + \frac{\partial L}{\partial \dot{q}_j}\frac{\mathrm{d}f_j}{\mathrm{d}t}\right) + \mathrm{O}(\epsilon^2) \tag{3.11}$$

and if the first-order term is zero, we say that L is invariant under the infinitesimal transformation specified by the functions f_i. I shall discuss the meaning of this shortly, but let us first derive its consequences. Using the equations of motion (3.5), and the fact that the coefficient of ϵ in (3.11) vanishes, we find that $\mathrm{d}F/\mathrm{d}t = 0$, where

$$F = \sum_i f_i \frac{\partial L}{\partial \dot{q}_i} = \sum_i f_i p_i \tag{3.12}$$

where p_i are the generalized momenta defined earlier. The quantity F is therefore conserved. For a classical system of point particles, this result constitutes what is known as *Noether's theorem.*

The simplest conservation law of this kind is the conservation of linear momentum, which follows from invariance under spatial translations. If we use Cartesian coordinates, a Lagrangian of the form (3.7) is unchanged when we replace each $\boldsymbol{x}_i$ by $\boldsymbol{x}_i + \epsilon\boldsymbol{a}$, where $\boldsymbol{a}$ is any constant vector, but the same for each particle. The velocities are unaffected because $\boldsymbol{a}$ is constant, and $\boldsymbol{a}$ cancels out of all the differences of pairs of coordinates. Thus, not only the first-order term but all the higher-order terms in (3.11) vanish. The conserved quantity F is $\boldsymbol{a} \cdot \boldsymbol{P}$, where $\boldsymbol{P} = \sum_i \boldsymbol{p}_i$ is the sum of the linear momenta of all the particles, or the total momentum of the system. So *if the Lagrangian is invariant under spatial translations, then the total linear momentum is conserved.* In the same way, invariance under rotations leads to the conservation of angular momentum, details of which are explored in exercise 3.1.

The symmetry transformations we have been using can be interpreted in two ways. According to what is known as the *active* point of view, by making the mathematical transformation $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{a}$, we are comparing the behaviour of the system when it occupies one or other of two regions of space, separated by the vector $\boldsymbol{a}$. Because the geometrical properties of our Galilean spacetime are the same everywhere, we expect that the laws of physics will be too. So the behaviour of the system, and therefore the form of the Lagrangian, should be the same in each location, so long as the system is isolated from any external influence. According to the *passive* point of view, we are comparing descriptions of the system referred to two sets of coordinates, whose origins are separated by the vector $-\boldsymbol{a}$. Again, since geometry is the same everywhere, equations of motion should have the same form, regardless of where we choose to place the origin of coordinates. Similar remarks apply to time translations and rotations.

Of course, these considerations apply to displacements or rotations of any size, not just infinitesimal ones. In fact, if the Lagrangian is unchanged at first order, it will also be unchanged by a large transformation which can be built from a sequence of infinitesimal ones. In general, however, it is only the infinitesimal ones which have the right form for the derivation to work. For example, the rotation $(x, y) \rightarrow (x \cos\epsilon + y \sin\epsilon, y \cos\epsilon - x \sin\epsilon)$ can be written, when ϵ is infinitesimal, as $(x, y) \rightarrow (x + \epsilon y, y - \epsilon x)$, and only the infinitesimal version can be used in (3.11). However, a rotation through a finite angle can obviously be built up from many infinitesimal ones. If the first-order change in L vanishes, then $x\partial L/\partial y = y\partial L/\partial x$, from which it is easy to show that L must be a function only of $(x^2 + y^2)$. But in that case, L is invariant under rotations through any angle.

3.3 The Hamiltonian

At the beginning of our discussion, we assumed that the state of a system would be uniquely specified by the coordinates and velocities of all its particles. For many theoretical purposes, however, the momenta play a more fundamental role than the velocities, and it is convenient to reformulate the theory in terms of them. To do this, we introduce a new function $H(\{q_i\}, \{p_i\})$ called the *Hamiltonian*. In terms of this function, a new set of equations of motion can be derived which are equivalent to the Euler–Lagrange equations, but which involve the momenta instead of the velocities.

The mathematical process of exchanging one set of variables for another is called a *Legendre transformation* and works as follows. We consider a set of small changes $\mathrm{d}q_i$ and $\mathrm{d}\dot{q}_i$ in the coordinates and velocities and write the corresponding small change in the Lagrangian as

$$\mathrm{d}L = \sum_i \left(\frac{\partial L}{\partial q_i} \mathrm{d}q_i + p_i \mathrm{d}\dot{q}_i \right) \tag{3.13}$$

where we have used the definition $p_i = \partial L/\partial \dot{q}_i$. Next, we define the Hamiltonian as

$$H(\{q_i\}, \{p_i\}) = \sum_i p_i \dot{q}_i - L \tag{3.14}$$

which implies that, on the right-hand side, all the velocities have been expressed in terms of the coordinates and momenta. Apart from this last step, the Hamiltonian is, of course, just the same as the total energy defined by (3.10). We can now use (3.13) to write down the small change in the Hamiltonian that results from a small change in the state of the system:

$$\begin{aligned} \mathrm{d}H &= \sum_i (p_i \mathrm{d}\dot{q}_i + \dot{q}_i \mathrm{d}p_i) - \mathrm{d}L \\ &= \sum_i \left(\dot{q}_i \mathrm{d}p_i - \frac{\partial L}{\partial q_i} \mathrm{d}q_i \right). \end{aligned} \tag{3.15}$$

According to the Euler–Lagrange equations (3.5), $\partial L/\partial q_i$ is equal to $\mathrm{d}p_i/\mathrm{d}t$. So, by allowing independent variations in each of the coordinates and momenta in turn, we may deduce from (3.15) the equations of motion

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \qquad\qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{3.16}$$

These are *Hamilton's equations.*

3.4 Poisson Brackets and Translation Operators

It may not be obvious that we have gained anything from these formal manipulations. In fact, when it comes to solving equations of motion for specific systems containing a few particles, it makes little practical difference whether we use the original equations of Newton, the Euler–Lagrange equations or Hamilton's equations: they all amount to the same thing, and exercise 3.2 invites readers to explore this equivalence in detail. However, the Lagrangian and Hamiltonian formulations of classical mechanics do reveal some mathematical features that are important for further developments. In modern theoretical physics, there are two situations in which an understanding of the mathematical structure of classical mechanics is especially useful. The first is that, when we deal with large collections of particles, it rapidly becomes impractical to solve the equations of motion directly. We must resort to a statistical description of such systems, and the Hamiltonian formulation is, as we shall discover in chapter 10, an indispensable tool for setting up this description.

An appreciation of the formal structure of classical mechanics is also useful when making the transition to quantum mechanics, which appears to supersede classical mechanics as a means of accounting for the behaviour of physical systems on atomic or sub-atomic scales. It is very difficult to infer directly from our experience what the rules of quantum mechanics should be. However, it turns out that the formal mathematical structures of classical and quantum mechanics have quite a lot in common. From a theoretical point of view, it seems to me that the most satisfactory way of approaching quantum theory is by exploiting the mathematical analogy with classical mechanics, which we shall explore in chapter 5. In this section, we shall construct some of the mathematical tools that make this analogy clear.

We saw in §3.2 that when the equations of motion are invariant under time translations, the total energy of the system, which is obtained by substituting into the Hamiltonian the actual coordinates and momenta of the particles, is conserved. Now, Hamilton's equations (3.16) offer us a deeper understanding of the role played by this quantity in the evolution of the state of the system with time. Suppose we wish to know how some quantity A changes with time, and that A can be expressed in terms of the coordinates and momenta as $A(\{q_i\}, \{p_i\})$. Using

Hamilton's equations, we can write

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \sum_i \left(\frac{\partial A}{\partial q_i}\dot{q}_i + \frac{\partial A}{\partial p_i}\dot{p}_i \right) = \{A, H\}_\mathrm{P} \tag{3.17}$$

where, for any two quantities A and B, the *Poisson bracket* $\{A, B\}_\mathrm{P}$ is defined as

$$\{A, B\}_\mathrm{P} = \sum_i \left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i} - \frac{\partial B}{\partial q_i}\frac{\partial A}{\partial p_i} \right). \tag{3.18}$$

It is implied, of course, that we treat the q_i and p_i as independent variables to evaluate the Poisson bracket and then substitute their actual values at time t to find the rate of change of A at that time.

Alternatively, we can define the differential operator $\mathcal{H}$ by

$$\mathcal{H} = i\{H, \quad\}_\mathrm{P} = \mathrm{i}\sum_i \left(\frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i} \right) \tag{3.19}$$

which means that $\mathcal{H}A = \mathrm{i}\{H, A\}_\mathrm{P} = -\mathrm{i}\{A, H\}_\mathrm{P}$ for any function A. The factor of i has no significance in classical mechanics, and I have included it just in order to bring out the quantum-mechanical analogy. Let us now make explicit the procedure for evaluating (3.17). We denote by $A(t)$ the value of A at time t, obtained by substituting into $A(\{q_i\}, \{p_i\})$ the functions $q_i(t)$ and $p_i(t)$ that describe the actual state of the system (they are solutions of Hamilton's equations). This substitution can be represented by using the Dirac delta function, which is described in appendix A for readers unfamiliar with its use. If we define

$$\rho(\{q_i\}, \{p_i\}, t) = \prod_i \delta\left(q_i - q_i(t)\right)\delta\left(p_i - p_i(t)\right) \tag{3.20}$$

then $A(t)$ can be written as

$$A(t) = \int \prod_i \mathrm{d}q_i \mathrm{d}p_i \, \rho\left(\{q\}, \{p\}, t\right) A\left(\{q\}, \{p\}\right). \tag{3.21}$$

To find $\mathrm{d}A/\mathrm{d}t$ from this expression, we can proceed in two ways. One is simply to differentiate, which gives $\partial\rho/\partial t$ inside the integral, since $A\left(\{q\}, \{p\}\right)$ does not depend on time. The other, according to (3.17), is to act on $A\left(\{q\}, \{p\}\right)$ with $i\mathcal{H}$. On integrating by parts, we see that this is equivalent to acting on ρ with $-\mathrm{i}\mathcal{H}$. The two results must be identical, so we find that ρ satisfies the equation

$$\mathrm{i}\frac{\partial\rho}{\partial t} = \mathcal{H}\rho \tag{3.22}$$

as readers may verify directly using (3.20), (3.19) and (3.17).

Readers who are familiar with elementary quantum mechanics will recognize (3.22) as having a similar form to Schrödinger's equation, and this was the main point of the exercise. Equation (3.17) can be written as $\mathrm{i}\mathrm{d}A/\mathrm{d}t = -\mathcal{H}A$, but it should be clear that this is not to be interpreted in quite the same way as (3.22). In (3.17), we use $\mathcal{H}$ to differentiate with respect to the q_i and p_i, treating them as independent variables, and then substitute the appropriate functions of time. On the other hand, ρ is a function of the q_i and p_i that appear as dummy integration variables in (3.21) and also of time, and (3.22) is to be taken at face value as a partial differential equation in all of these variables. Bearing these points in mind, we can express $A(t)$ as a Taylor series

$$\begin{aligned} A(t) &= \sum_{n=0}^{\infty} \frac{1}{n!} t^n A^{(n)}(0) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} (\mathrm{i}t\mathcal{H})^n A \\ &= \exp\{\mathrm{i}t\mathcal{H}\}A. \end{aligned} \tag{3.23}$$

Here, the nth derivative of $A(t)$ evaluated at $t = 0$ is denoted by $A^{(n)}(0)$, and the derivative can be replaced by $\mathrm{i}\mathcal{H}$ in the manner I have just described. The exponential of the differential operator is a convenient shorthand for the power series. Obviously, we evaluate the final expression by substituting the $q_i(0)$ and $p_i(0)$ corresponding to the state at $t = 0$ after acting with $\mathcal{H}$. The exponential operator is responsible for transforming $A(0)$ into $A(t)$ and in this context $\mathcal{H}$ is called the *generator of time translations*.

In Cartesian coordinates, we can transform any function $f(\{\boldsymbol{x}_i\})$ of the coordinates into $f(\{\boldsymbol{x}_i + \boldsymbol{a}\})$ by means of a similar Taylor series using the operator $\exp\{\mathrm{i}\boldsymbol{a} \cdot \boldsymbol{\mathcal{P}}\}$, where the *generator of spatial translations* is

$$\boldsymbol{\mathcal{P}} = -\mathrm{i}\sum_i \nabla_i. \tag{3.24}$$

The sum here is over the N particles in the system rather than the $3N$ coordinates. It is easy to see that this generator may be written in a form similar to (3.19) as $\boldsymbol{\mathcal{P}} = \mathrm{i}\{\boldsymbol{P},\quad\}_{\mathrm{P}}$, where $\boldsymbol{P}$ is the total linear momentum, and we recall that $\boldsymbol{P}$ is the quantity whose conservation law follows from invariance under spatial translations. Again, knowledgeable readers will recognize (3.24) as being closely related to the momentum operator that acts on quantum-mechanical wavefunctions.

Equation (3.22) also serves as the starting point of classical statistical mechanics, if we regard ρ as expressing the probability that the coordinates and momenta have, at time t, the values $\{q_i\}$ and $\{p_i\}$. Then (3.21) is the usual expression for the mean value of A. In the case we have considered, the probability is zero unless the coordinates and momenta correspond to the evolution of the system from a definite initial state, but more general probability distributions can be constructed, as we shall see in chapter 10. In this context, (3.22) is called the *Liouville equation* and $\mathcal{H}$ the *Liouville operator*.

3.5 The Action Principle in Minkowski Spacetime

In earlier sections of this chapter, we have investigated the way in which the geometrical structure of Galilean spacetime constrains the possible kinds of behaviour of particles that live there. A source of difficulty was the fact that the geometrical roles of space and time are quite different. This leads to a certain amount of confusion about the exact significance of invariance under Galilean transformations and the meaning of inertial frames of reference. In particular, it does not seem to be possible to arrive at a purely geometrical definition of inertial frames that is independent of considerations about the way in which physical objects are actually observed to behave. In the Minkowski spacetime of special relativity, and in the more general spacetimes envisaged in general relativity and similar theories, space and time appear on much the same footing, and a more clear-cut discussion is possible. Conversely, to my mind, the relativistic view makes it rather more difficult to understand the obvious dissimilarity of space and time as they enter our conscious experience. I do not propose to enter into the philosophical perplexities of this question here, but interested readers may like to consult, for example, the books by Block *et al* (1997), Landsberg (1982), Lockwood (1989), Lucas (1973), Morris (1986), Ornstein (1969), Prigogine (1980), Smart (1964) and Whitrow (1975).

We learned in chapter 2 that the relativistic spacetimes are manifolds whose points can be labelled by a set of four coordinates x^μ ($\mu = 0, 1, 2, 3$). The separation of two points cannot be uniquely decomposed into spatial and temporal components. What we can do is to assign a *proper time interval* to a specific curve that joins them. The proper time interval $\mathrm{d}\tau$ for an infinitesimal segment of the curve is given by (2.7). In that expression, the coefficients $g_{\mu\nu}$ are the components of the metric tensor, which contains all our information about the geometrical structure. In general, they vary from point to point and their values depend on the coordinate system we are using. The value of $\mathrm{d}\tau$ is the same in all coordinate systems, however. If the metric tensor is that of Minkowski spacetime then, by definition, it will be possible to find a Cartesian coordinate system (and, in fact, infinitely many of them) such that its components are given by the matrix (2.8). Relative to such a system, time is measured by x^0/c, where c is the speed of light, while the other three coordinates measure spatial distances.

We may now define an *inertial system of Cartesian coordinates* as one in which the metric tensor has the special form (2.8). More generally, an inertial system is one that can be obtained from an inertial Cartesian system by keeping the time coordinate and redefining the spatial ones in a time-independent manner. For example, if we simply exchange (x^1, x^2, x^3) for polar coordinates (r, θ, ϕ)

we still have an inertial system, but if we exchange them for a set of rotating axes, we get a non-inertial system. In the rest of this chapter, I shall use only inertial Cartesian coordinates.

As with Galilean spacetime, we want to see how geometrical symmetries constrain the behaviour of physical systems. These symmetries consist of all the coordinate transformations that leave the form of the metric tensor unchanged: that is, they convert one inertial frame into another. They are called *isometries*, meaning ‘same metric’. Space and time translations can now be considered together. They are transformations of the type $x^{\mu'} = x^\mu + a^\mu$, where a^μ are the components of a constant 4-vector. We see from (2.6) that this leaves $g_{\mu\nu}$ unchanged, since $\mathrm{d}x^{\mu'} = \mathrm{d}x^\mu$. The other isometries are *Lorentz transformations*. These include both spatial rotations and ‘boosts’, such as (2.2), which relate two systems with a constant relative velocity. They can be expressed in the form

$$x^{\mu'} = \Lambda^{\mu'}{}_{\mu} x^\mu \tag{3.25}$$

where, as in chapter 2, we are using a prime on the index μ to indicate the new coordinates. For example, a rotation about the x^1 axis through an angle θ corresponds to the transformation matrix

$$\Lambda^{\mu'}{}_{\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{pmatrix} \tag{3.26}$$

while the boost written in (2.2) is represented by

$$\Lambda^{\mu'}{}_{\mu} = \begin{pmatrix} \cosh\alpha & -\sinh\alpha & 0 & 0 \\ -\sinh\alpha & \cosh\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.27}$$

with $\sinh\alpha = (1 - v^2/c^2)^{-1/2} v/c$ (and so $\cosh\alpha = (1 - v^2/c^2)^{-1/2}$). The set of all rotations and boosts is called the *proper Lorentz group*. The set of all rotations, boosts and translations is called the *proper Poincaré group*. The full *Poincaré group* includes time reversal and space reflections, $(x^{0'}, x^{1'}, x^{2'}, x^{3'}) = (-x^0, x^1, x^2, x^2)$ or $(x^0, -x^1, x^2, x^3)$, etc, and is the *isometry group* of Minkowski spacetime.

Any Poincaré transformation—that is, the net effect of any sequence of translations, rotations and boosts—can be expressed as $x^{\mu'} = \Lambda^{\mu'}{}_{\mu} x^\mu + a^{\mu'}$. Let $f(x)$ be a scalar function of the coordinates (one that depends on the spacetime point, but not on the choice of coordinate system). Under a Poincaré transformation, both infinitesimal coordinate differences and derivatives of scalar functions transform in a manner that depends only on Λ:

$$\mathrm{d}x^{\mu'} = \Lambda^{\mu'}{}_{\mu} \mathrm{d}x^\mu \tag{3.28}$$

$$\partial_{\mu'} f = \Lambda^{\mu}{}_{\mu'} \partial_\mu f. \tag{3.29}$$

(Recall the following from chapter 2: repeated indices, occurring once in the upper position and once in the lower position are summed over; ∂_μ is an abbreviation for $\partial/\partial x^\mu$; the matrix $\Lambda^{\mu}{}_{\mu'}$ is the inverse of $\Lambda^{\mu'}{}_{\mu}$—see (2.14).) An object with four components V^μ that transform like $\mathrm{d}x^\mu$ is called a *contravariant 4-vector*; an object with components V_μ that transforms like $\partial_\mu f$ is a *covariant 4-vector*. More complicated entities, with transformation laws similar to (2.19), are *4-tensors*: for example, the metric tensor, with two lower indices, is said to have covariant rank 2. These 4-tensors are not necessarily true tensors as defined in chapter 2, because we are considering only Λ matrices with constant elements. For example, $\partial_\mu V^\nu$ is a 4-tensor, but not a true tensor. Readers may readily verify that any expression such as $\eta_{\mu\nu}U^\mu U^\nu$ composed of tensors, in which all indices appear in pairs and the implied summations have been carried out (the process called *contraction* in chapter 2), is invariant under Lorentz transformations: it is a Lorentz scalar.

The path of a particle through Minkowski spacetime may be described parametrically by a set of four functions $x^\mu(\tau)$, each point on the path being labelled by a value of the proper time τ. Since τ is a scalar, the set of functions $\mathrm{d}x^\mu/\mathrm{d}\tau$ are the components of a 4-vector, the tangent vector to the path. As in our discussion of Galilean spacetime, we expect the equations of motion for an isolated system to have the same form in any two coordinate systems in which the metric tensor is the same. Thus, the form of these equations should be unchanged by any Poincaré transformation: we say that they should be *covariant* under these transformations. To achieve this, we need an action which is Poincaré invariant. That is, the action must be a Lorentz scalar and translationally invariant. Following the arguments of §3.1, we see that for a single particle it must be of the form

$$S = \int \mathrm{d}\tau\, L(\eta_{\mu\nu}\dot{x}^\mu\dot{x}^\nu) \tag{3.30}$$

where $\dot{x}^\mu$ denotes $\mathrm{d}x^\mu/\mathrm{d}\tau$. Using the notation $X = \frac{1}{2}\eta_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$, we find that the Euler–Lagrange equations are

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2}\frac{\mathrm{d}L}{\mathrm{d}X} + \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}X}{\mathrm{d}\tau}\frac{\mathrm{d}^2L}{\mathrm{d}X^2} = 0. \tag{3.31}$$

In Galilean spacetime, the function $L(X)$ could be determined by requiring invariance under Galilean transformations. Here, this symmetry is replaced by Lorentz invariance, which we have already taken into account. In fact, the form of $L(X)$ is quite irrelevant! According to (2.6), when $x^\mu(\tau)$ is the actual path of a particle through Minkowski spacetime, it must satisfy $X = \frac{1}{2}c^2$ and therefore $\mathrm{d}X/\mathrm{d}\tau = 0$ as well as (3.31). Therefore, the only feature of L that has any real meaning is the value of $\mathrm{d}L/\mathrm{d}X$ at $X = \frac{1}{2}c^2$. As long as this value is non-zero, the equation of motion is simply $\mathrm{d}^2x^\mu/\mathrm{d}\tau^2 = 0$. We may as well make the simplest choice

$$L = -\tfrac{1}{2}m\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} \tag{3.32}$$

where, as before, m will be identified with the mass of the particle. (Many authors refer to m as the 'rest mass' to distinguish it from a velocity-dependent 'mass', which is in fact the energy divided by c^2. I do not recommend this practice and will not follow it in this book.) In a frame of reference where the particle moves very slowly compared to c, the proper time τ is approximately equal to t, and $x^0 = ct$. In this frame, therefore, we find $L \approx -\frac{1}{2}m(c^2 - \dot{\boldsymbol{x}}^2)$, which differs only by an unimportant constant $-\frac{1}{2}mc^2$ from the Lagrangian for a Newtonian particle.

The canonical momenta obtained from this Lagrangian, which are conserved as a consequence of translational invariance, are the four components of the *energy–momentum 4-vector* or *4-momentum*

$$p_\mu = -\frac{\partial L}{\partial \dot{x}^\mu} = m\eta_{\mu\nu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} \tag{3.33}$$

or in the contravariant form $p^\mu = \eta^{\mu\nu}p_\nu = m\mathrm{d}x^\mu/\mathrm{d}\tau$. (The contravariant version of the metric tensor $\eta^{\mu\nu}$ used here to 'raise' the index is the inverse of the matrix (2.8), which is numerically the same matrix, as long as we confine ourselves to Cartesian coordinates.) This definition differs by a minus sign from the one that we used in the Galilean theory. The sign results from my convention about the sign of $\eta_{\mu\nu}$ (see section 2.4) and is needed to make the contravariant momentum p^μ agree with what we normally call energy and momentum. (The mathematics would work perfectly well with either sign, so long as we do things consistently.) The velocity of the particle relative to the frame of reference with coordinates (ct, x^1, x^2, x^3) is $\boldsymbol{u} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$. We see from (2.3) that $\mathrm{d}\tau/\mathrm{d}t = (1 - u^2/c^2)^{1/2}$, so using $\mathrm{d}x^\mu/\mathrm{d}\tau = (\mathrm{d}\tau/\mathrm{d}t)^{-1}\mathrm{d}x^\mu/\mathrm{d}t$, we can write the 4-momentum as

$$(p^0, \boldsymbol{p}) = \left(\frac{mc}{(1 - u^2/c^2)^{1/2}}, \frac{m\boldsymbol{u}}{(1 - u^2/c^2)^{1/2}}\right). \tag{3.34}$$

Since this is conserved, we may identify the zeroth, time-like component as $1/c$ times the energy (to make its dimensions agree with the non-relativistic definition) and the other three as the linear momentum. Using either (3.34) or (3.33) we find that $p_\mu p^\mu = m^2\eta_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}\tau)(\mathrm{d}x^\nu/\mathrm{d}\tau) = m^2c^2$.

Because there is no unique time in Minkowski spacetime, the integration variable τ in (3.30) is associated with the path of a specific particle. The action for a collection of non-interacting particles, labelled by i, following paths $x_i^\mu(\tau_i)$ is therefore

$$S = -\sum_i \int \mathrm{d}\tau_i \, \tfrac{1}{2}m_i\eta_{\mu\nu}\frac{\mathrm{d}x_i^\mu}{\mathrm{d}\tau_i}\frac{\mathrm{d}x_i^\nu}{\mathrm{d}\tau_i}. \tag{3.35}$$

It will soon be useful to us to have expressions for the number density $n(x)$ (number per unit volume) and current density $\boldsymbol{j}(x)$ (number crossing unit area per unit time) of these particles. At the microscopic level, these are zero unless the point x lies exactly on the path of one of the particles. They may be written as

$$n(t,\boldsymbol{x}) = \sum_i \delta^3\left(\boldsymbol{x} - \boldsymbol{x}_i(t)\right) \tag{3.36}$$

$$\boldsymbol{j}(t,\boldsymbol{x}) = \sum_i \frac{\mathrm{d}\boldsymbol{x}_i(t)}{\mathrm{d}t}\delta^3\left(\boldsymbol{x} - \boldsymbol{x}_i(t)\right). \tag{3.37}$$

So long as no particles are created or destroyed, they should satisfy the equation of continuity $\partial n/\partial t + \nabla \cdot \boldsymbol{j} = 0$. Readers are invited to verify this and to consider what happens if particles *are* created or destroyed. Using the fact that $\mathrm{d}x^0/\mathrm{d}t = c$, we can assemble the quantities (3.36) and (3.37) into a 4-vector

$$j^\mu(t,\boldsymbol{x}) = \left(cn(t,\boldsymbol{x}),\, j^1(t,\boldsymbol{x}),\, j^2(t,\boldsymbol{x}),\, j^3(t,\boldsymbol{x})\right) = \sum_i \frac{\mathrm{d}x_i^\mu(t)}{\mathrm{d}t}\,\delta^3\left(\boldsymbol{x} - \boldsymbol{x}_i(t)\right). \tag{3.38}$$

Although $\mathrm{d}x^\mu$ is a 4-vector, neither $\mathrm{d}t$ nor the δ function is a scalar, so it is not obvious that this really is a 4-vector, which would transform correctly under Lorentz transformations. It is left as an exercise for readers to show that the 4-vector current density can be rewritten in the form

$$j^\mu(x) = c\sum_i \int \mathrm{d}\tau_i \frac{\mathrm{d}x^\mu(\tau_i)}{\mathrm{d}\tau_i}\,\delta^4\left(x - x_i(\tau_i)\right) \tag{3.39}$$

which manifestly *is* a 4-vector. In terms of j^μ, the equation of continuity reads

$$\partial_\mu j^\mu = 0. \tag{3.40}$$

A current that satisfies this equation is said to be a *conserved current.*

If A is some physical quantity carried by the particles, we can define a current whose zeroth component is the density of A (the amount of A per unit volume) and whose spatial components represent the rate at which A is transported by the flow of particles (the amount of A carried across unit area per unit time). It is

$$j_A^\mu(x) = c\sum_i \int \mathrm{d}\tau_i\; A_i \frac{\mathrm{d}x^\mu(\tau_i)}{\mathrm{d}\tau_i}\,\delta^4\left(x - x_i(\tau_i)\right) \tag{3.41}$$

where A_i is the amount of A carried by the ith particle. Two important examples are the *electromagnetic current*, obtained by taking A to be electric charge, and the *stress–energy–momentum tensor*, which I shall refer to as the *stress tensor* for brevity. This tensor is formed from the four currents obtained by taking A to be the components of the 4-momentum:

$$T^{\mu\nu}(x) = c\sum_i \int \mathrm{d}\tau_i\, m_i \frac{\mathrm{d}x^\mu(\tau_i)}{\mathrm{d}\tau_i}\frac{\mathrm{d}x^\nu(\tau_i)}{\mathrm{d}\tau_i}\,\delta^4\left(x - x_i(\tau_i)\right). \tag{3.42}$$

The stress tensor plays a central role in the relativistic theory of gravity. It is symmetric in the indices μ and ν and is conserved, since $\partial_\nu T^{\mu\nu} = 0$, as readers are invited to prove. This simply reflects the fact that energy and momentum are conserved quantities, so their densities and currents must obey the equation of continuity. It should be borne in mind, however, that (3.42) is the stress tensor for a collection of non-interacting particles. If, for example, the particles interact via electromagnetic fields, then energy and momentum can be transferred to and from these fields and the stress tensor will be conserved only when a suitable electromagnetic contribution is included. The same goes for fields associated with other forces, including gravitational fields, but the nature of conservation laws in non-Minkowski spacetimes can be a little subtle.

A simple example of a stress tensor is afforded by what cosmologists call a *perfect fluid.* This is a fluid that has a rest frame, in which its density is spatially uniform and the average velocity of its particles is zero. For such a fluid, as discussed in exercise 3.4, the stress tensor is

$$T^{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} \tag{3.43}$$

where ρ is the energy density and p the pressure.

3.6 Classical Electrodynamics

The only fully-fledged classical theory of interacting particles in Minkowski spacetime is electrodynamics, in which the forces are described by electric and magnetic fields $\boldsymbol{E}(t, \boldsymbol{x})$ and $\boldsymbol{B}(t, \boldsymbol{x})$, which obey *Maxwell's equations.* In a suitable system of units, these equations are

$$\nabla \cdot \boldsymbol{E} = \rho_\mathrm{e} \tag{3.44}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{3.45}$$

$$\nabla \times \boldsymbol{E} + \frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = 0 \tag{3.46}$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \frac{1}{c}\boldsymbol{j}_\mathrm{e} \tag{3.47}$$

where ρ_e is the electric charge density and $\boldsymbol{j}_\mathrm{e}$ is the electric current density. The first of these equations is Gauss' law which, for a static charge distribution, is a simple consequence of the Coulomb force law. The second asserts that there are no magnetic monopoles, which would be the magnetic analogues of electric charges. The *grand unified theories* of fundamental forces discussed in chapter 12 suggest that such monopoles may exist but, at the time of writing, there is no firm evidence that they do. The third equation (3.46) is Faraday's law, which describes the generation of electric fields by time-varying magnetic fields, and the fourth (3.47) is Ampère's law which, conversely, describes the generation of magnetic fields by both the flow of electric currents and changing electric fields. Readers who are not familiar with the derivation of these equations from simple physical observations will find this discussed in any standard textbook on electromagnetic theory. This form of Maxwell's equations is valid in the Heaviside–Lorentz system of units and is the microscopic version. The fields $\boldsymbol{D}$ and $\boldsymbol{H}$ that are often used to take approximate account of the properties of dielectric and magnetic materials on a macroscopic scale are not used here.

As far as the classical theory is concerned, I know of no convincing way of arriving at Maxwell's equations other than by inferring them from experimental observations. On the other hand, we shall see in chapter 8 that in quantum mechanics they arise in a rather natural way from geometrical considerations. For now, we shall take them as given and briefly derive some important and elegant properties. Two of the equations, (3.45) and (3.46), are satisfied automatically if we express the fields in terms of an electric scalar potential $\phi(t, \boldsymbol{x})$ and a magnetic vector potential $\boldsymbol{A}(t, \boldsymbol{x})$ as

$$\boldsymbol{E} = -\nabla\phi - \frac{1}{c}\frac{\partial \boldsymbol{A}}{\partial t} \tag{3.48}$$

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \tag{3.49}$$

which follows from the identities $\nabla \times \nabla\phi \equiv 0$ and $\nabla \cdot (\nabla \times \boldsymbol{A}) \equiv 0$. The two remaining equations take on a much more compact appearance if we express them in 4-vector notation. The potentials can be assembled into a contravariant 4-vector A^μ with components $(\phi, \boldsymbol{A})$ or its covariant version A_μ with components $(\phi, -\boldsymbol{A})$. The electric and magnetic fields then form the components of an antisymmetric *field strength 4-tensor*

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{3.50}$$

whose contravariant form may be written explicitly as

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E^1 & -E^2 & -E^3 \\ E^1 & 0 & -B^3 & B^2 \\ E^2 & B^3 & 0 & -B^1 \\ E^3 & -B^2 & B^1 & 0 \end{pmatrix}. \tag{3.51}$$

In terms of this tensor, the remaining Maxwell equations (3.44) and (3.47) are simply

$$\partial_\mu F^{\mu\nu} = \frac{1}{c} j_\mathrm{e}^\nu \tag{3.52}$$

where j_e^ν is the 4-vector current density with components $(c\rho_\mathrm{e}, \boldsymbol{j}_\mathrm{e})$.

These equations can be derived from an action principle in more or less the same way as the equations of motion for particles. Because we are now dealing with electromagnetic fields that exist at each point of spacetime rather than with the trajectories of particles, the action must be written as the integral over all space and time of a *Lagrangian density* $\mathcal{L}$:

$$S = \frac{1}{c}\int \mathrm{d}^4x\mathcal{L}(x) \tag{3.53}$$

where

$$\mathcal{L}(x) = -\frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x) - \frac{1}{c}j_{\mathrm{e}}^{\mu}(x)A_{\mu}(x). \tag{3.54}$$

The factor $1/c$ in (3.53) arises from the fact that $x^0 = ct$. By varying A_μ, readers may readily verify that the Euler–Lagrange equations are (3.52). To obtain a complete theory of charged particles, we must add to (3.53) the action (3.35) for the particles themselves.

Consider the case of a single particle with charge q. The current is given by (3.41) with $A = q$ and, on substituting this into (3.53), the spacetime integral in the $j_{\mathrm{e}}^{\mu}A_\mu$ term can be carried out. Thus, the total action is given by

$$\begin{aligned} S = &-\int \mathrm{d}\tau \frac{1}{2}m\eta_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} - \frac{q}{c}\int \mathrm{d}\tau\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}A_\mu\big(x(\tau)\big) \\ &- \frac{1}{4c}\int \mathrm{d}^4x\, F_{\mu\nu}F^{\mu\nu}. \end{aligned} \tag{3.55}$$

By varying the path of the particle, we find the equation of motion

$$m\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} = \frac{q}{c}\eta_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}F^{\mu\sigma}. \tag{3.56}$$

Its zeroth component can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{mc^2}{(1-u^2/c^2)^{1/2}}\right) = q\boldsymbol{u}\cdot\boldsymbol{E} \tag{3.57}$$

which asserts that the rate of change of the energy of the particle is the rate at which work is done on it by the electric field, while the spatial components reproduce the usual Lorentz force

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = q\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{u}\times\boldsymbol{B}\right). \tag{3.58}$$

The momentum $\boldsymbol{p}$ here is that written in (3.34). However, the components of the 4-momentum shown there are now not equal to the canonical momenta conjugate to the coordinates of the particle, which are

$$p^\mu_{\mathrm{can}} = m\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} + \frac{q}{c}A^\mu\big(x(\tau)\big). \tag{3.59}$$

The canonical structure of electrodynamics is explored further in the exercises.

Electromagnetism possesses an important symmetry known as *gauge invariance*. In the classical theory, this symmetry seems to appear more or less by accident but, as we shall see in chapter 8, it has a deep-seated significance in quantum mechanics and underlies most of our present understanding of the fundamental forces of nature. Let $\theta(x)$ be any function of x and consider redefining the 4-vector potential according to

$$A'_\mu(x) = A_\mu(x) - \partial_\mu \theta(x). \tag{3.60}$$

The field strengths given by (3.50) are the same functions of A'_μ as they were of A_μ, because the $\partial_\mu \partial_\nu \theta$ terms cancel. This clearly has to do with the antisymmetry of $F_{\mu\nu}$. This antisymmetry also has the consequence that the electric current must be conserved (it must obey (3.40)), as we see by differentiating (3.52). Suppose we demand that the action (3.53) with Lagrangian density (3.54) should be *gauge invariant*: that is, its form should be preserved after the change of variable (3.60), which is called a *gauge transformation*. The change in the action is $-(1/c)\int \mathrm{d}^4x\, j_\mathrm{e}^\mu\, \partial_\mu \theta$ so, after integrating by parts, we see that this vanishes provided that the current is conserved. Therefore, the quantity whose conservation is associated with the symmetry of gauge invariance is electric charge. If there is no mechanism whereby charged particles can be created or destroyed, then electric charge will naturally be conserved. If there is such a mechanism, then charge may or may not be conserved and, if it is not, then the presence of electromagnetic forces will not make it so. In the latter case, (3.52) could not be true, and Maxwell's theory would not be self-consistent. Readers will recall (I hope!) that the so-called *displacement current* $\partial \boldsymbol{E}/\partial t$ in (3.47) was introduced by Maxwell precisely in order to make his equations consistent with the conservation of electric charge. Experimentally, of course, even though individual charged particles can be created and destroyed, these processes are always found to occur in such a way that electric charge is conserved overall.

3.7 Geometry in Classical Physics

This section is something of a detour from our main line of enquiry. Its purpose is to offer a glimpse of the geometrical view of classical physics that is often encountered in the more advanced literature and of some of the associated terminology. We shall see, in particular, how Maxwell's equations can be expressed in an extremely compact form, once we have the appropriate geometrical tools to hand, and that the Poisson bracket (3.18), which we met in connection with Hamilton's equations, can be understood as part of a geometrical structure that captures the essence of classical mechanics in a rather elegant manner. The perspective we shall gain serves to illustrate the remarkable unifying power of modern differential geometry as applied to theoretical physics (which extends in important ways to the study of quantum as well as classical phenomena). On the other hand, we shall learn no essentially new physics and the remainder of the Tour will not make extensive use of the new geometrical tools, so this section might well be omitted at a first reading.

3.7.1 More on tensors

In §2.2, I showed how vectors and one-forms can be defined as geometrical objects in their own right, but then took the easy option of defining higher-rank tensors in terms of the transformation laws for their components, referred to definite systems of coordinates. It will now be useful to see how these higher-rank tensors can be defined without recourse to coordinates. Recall that a one-form is a linear function whose argument is a vector and whose value is a scalar. A rank $\binom{0}{n}$ tensor $\boldsymbol{T}$ can be defined similarly as a multilinear, scalar-valued function of n arguments, each of which is a vector. If we *do* use components, then the value of this function is

$$\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{V}, \cdots) = T_{ab\ldots} U^a V^b \cdots . \tag{3.61}$$

Here, I use Latin indices $a, b, \ldots$ (each of which has values $1, 2, \ldots, d$) to indicate coordinates in a general d-dimensional manifold, reserving $\mu, \nu, \ldots$ (with values running from 0 to $d - 1$) to indicate that the manifold is a relativistic spacetime. The term 'multilinear' means that $\boldsymbol{T}$ is a linear function of each of its arguments:

$$\boldsymbol{T}(\boldsymbol{U}, \alpha\boldsymbol{V} + \beta\boldsymbol{W}, \cdots) = \alpha\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{V}, \cdots) + \beta\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{W}, \cdots) \tag{3.62}$$

and similarly for all the other arguments. Unless the tensor has a special symmetry, the order of the arguments is important. That is to say, $\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{V}, \cdots)$ does not necessarily mean the same as $\boldsymbol{T}(\boldsymbol{V}, \boldsymbol{U}, \cdots)$. Furthermore, a rank $\binom{m}{n}$ tensor is a multilinear, scalar-valued function of $m + n$ arguments, of which n are vectors and m are one-forms. In components, we have, for example

$$\boldsymbol{T}(\boldsymbol{U}, \omega, \boldsymbol{V}) = T_a{}^b{}_c U^a \omega_b V^c . \tag{3.63}$$

Since a vector is a rank $\binom{1}{0}$ tensor, this definition tells us that it is a linear function, whose value is a scalar and whose argument is a one-form. Originally, of course, we defined a vector as a differential operator $\mathrm{d}/\mathrm{d}\lambda$ representing a rate of change along a curve parametrized by λ. Readers who have difficulty in reconciling these two points of view, or who suspect an element of circularity in this entire sequence of definitions, may find it helpful to reflect on the example of a one-form ω_f, which represents the gradient of a scalar field f. To say that ω_f is a function of vectors means that we have a specific scalar field f whose gradient is ω_f and, given any curve with tangent vector $\boldsymbol{V}$, we can find the rate of change of f (namely $\omega_f(\boldsymbol{V}) = \mathrm{d}f/\mathrm{d}\lambda$) along this curve. To say that $\boldsymbol{V}$ is a function of one-forms means that we have a specific curve whose tangent vector is $\boldsymbol{V}$ and, given any scalar field f with gradient ω_f, we can find its rate of change along our curve (namely $\boldsymbol{V}(\omega_f) = \mathrm{d}f/\mathrm{d}\lambda$). In terms of components, the symmetry of the expressions $\omega(\boldsymbol{V}) = \omega_a V^a = \boldsymbol{V}(\omega)$ makes the equivalence of these two points of view rather obvious.

Given a system of coordinates x^a, we saw in (2.11) that the partial derivatives $\partial/\partial x^a$ serve as a set of basis vectors. Correspondingly, we can introduce a set of basis one-forms, which are denoted by $\mathrm{d}x^a$ and specified by giving their values when presented with any basis vector as an argument:

$$\mathrm{d}x^a(\partial/\partial x^b) = \delta^a_b. \tag{3.64}$$

To my physicist's eye, this notation is a little disconcerting. In particular, we must be careful not to confuse the one-form $\mathrm{d}x^a$ with an infinitesimal coordinate difference $\mathrm{d}x^a$, which looks exactly the same but is actually a component of a vector! It is worth noting, though, that these two different objects transform in the same way under a change of coordinates. In fact, a one-form $\omega = \omega_a \mathrm{d}x^a$ is a coordinate-independent object, so we must have $\mathrm{d}x^{a'} = \Lambda^{a'}{}_a \mathrm{d}x^a$, in order that

$$\omega = \omega_{a'}\mathrm{d}x^{a'} = \omega_a \Lambda^a{}_{a'}\Lambda^{a'}{}_b \mathrm{d}x^b = \omega_a\, \delta^a_b\, \mathrm{d}x^b = \omega_a \mathrm{d}x^a. \tag{3.65}$$

Thus, basis one-forms transform in the same way as the components of a vector. Evidently, the converse is also true: basis vectors $\partial/\partial x^a$ transform in the same way as the components of a one-form, such as $\partial f/\partial x^a$.

Bases for tensors of higher rank can be constructed by means of the *tensor product*, $\otimes$, which is defined as follows. Suppose that $\boldsymbol{S}$ is a rank $\binom{m}{n}$ tensor and $\boldsymbol{T}$ is a rank $\binom{m'}{n'}$ tensor. Then $\boldsymbol{S} \otimes \boldsymbol{T}$ is the rank $\binom{m+m'}{n+n'}$ tensor such that

$$\boldsymbol{S} \otimes \boldsymbol{T}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{m+n}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m'+n'}) = \boldsymbol{S}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{m+n})\boldsymbol{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m'+n'}) \tag{3.66}$$

where each of the arguments $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ is either a vector or a one-form, as required by the character of $\boldsymbol{S}$ and $\boldsymbol{T}$. The right-hand side is just the ordinary product of two numbers (or, in the case of tensor fields, of two scalar fields) $\boldsymbol{S}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{m+n})$ and $\boldsymbol{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m'+n'})$ and the components of $\boldsymbol{S} \otimes \boldsymbol{T}$ are the ordinary products

$$(\boldsymbol{S} \otimes \boldsymbol{T})^{ab\ldots ef\ldots}_{cd\ldots gh\ldots} = S^{ab\ldots}_{cd\ldots}\, T^{ef\ldots}_{gh\ldots}. \tag{3.67}$$

In particular, the product $\mathrm{d}x^a \otimes \mathrm{d}x^b \otimes \mathrm{d}x^c \cdots$ is the covariant tensor which, when presented with the vector arguments $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$, ... *in that order*, produces the value

$$\mathrm{d}x^a(\boldsymbol{U})\mathrm{d}x^b(\boldsymbol{V})\mathrm{d}x^c(\boldsymbol{W}) \cdots = U^a V^b W^c \cdots. \tag{3.68}$$

It should now be readily understood that a wholly covariant tensor, say of rank $\binom{0}{n}$, can be expressed as a linear combination

$$\boldsymbol{T} = T_{a_1 a_2 \ldots a_n} \mathrm{d}x^{a_1} \otimes \mathrm{d}x^{a_2} \cdots \otimes \mathrm{d}x^{a_n} \tag{3.69}$$

and that other tensors can be expressed as linear combinations of appropriate tensor products of basis one-forms and basis vectors.

3.7.2 Differential forms, dual tensors and Maxwell's equations

Astute readers will long ago have suspected that where there are one-forms, there ought also to be 2-forms, 3-forms and so on. Indeed there are. A *2-form* is an antisymmetric rank $\binom{0}{2}$ tensor. In coordinate-free language, this means that $\omega(\boldsymbol{U}, \boldsymbol{V}) = -\omega(\boldsymbol{V}, \boldsymbol{U})$ for any two vectors $\boldsymbol{U}$ and $\boldsymbol{V}$; in terms of components it means that $\omega_{ab} = -\omega_{ba}$. A *p-form* is a totally antisymmetric rank $\binom{0}{p}$ tensor. That is, it changes sign when *any* two neighbouring arguments or indices are interchanged: $\omega(\boldsymbol{U}, \ldots, \boldsymbol{V}, \boldsymbol{W}, \ldots) = -\omega(\boldsymbol{U}, \ldots, \boldsymbol{W}, \boldsymbol{V}, \ldots)$ or $\omega_{a \ldots bc \ldots} = -\omega_{a \ldots cb \ldots}$. As a matter of fact, the tensor also changes sign when two non-neighbouring arguments or indices are interchanged, $\omega_{a \ldots b \ldots c \ldots} = -\omega_{a \ldots c \ldots b \ldots}$, because moving b and c to their new positions one step at a time always requires an odd number of steps in total. In component language, it should be clear that $\omega_{ab \ldots} = 0$ if any two indices are equal. In a d-dimensional manifold, each index can take only d different values, so if there are more than d indices, at least two of them must be the same. Thus, p-forms with $p > d$ do not exist (or, at least, they are uninteresting, being identically zero). For $p = d$, the component $\omega_{a_1 a_2 \ldots a_d}$ vanishes unless its indices $(a_1, a_2, \ldots, a_d)$ have values that are a permutation of $(1, 2, \ldots, d)$, in which case it is equal to $\pm\omega_{12 \ldots d}$. Every d-form is therefore proportional to the Levi-Civita tensor density $\epsilon_{a_1 a_2 \ldots a_d}$ (discussed in appendix A for the case $d = 4$) whose components are 1 for an even permutation, -1 for an odd permutation and zero otherwise.

A basis for 2-forms is constructed by defining the *wedge product*

$$\omega \wedge \sigma \equiv \omega \otimes \sigma - \sigma \otimes \omega \tag{3.70}$$

for any two 1-forms ω and σ. The object $\omega \wedge \sigma$ is a 2-form, because its value when presented with two vector arguments $\boldsymbol{U}$ and $\boldsymbol{V}$ *in that order* is

$$\omega \wedge \sigma(\boldsymbol{U}, \boldsymbol{V}) = \omega(\boldsymbol{U})\sigma(\boldsymbol{V}) - \omega(\boldsymbol{V})\sigma(\boldsymbol{U}). \tag{3.71}$$

Clearly, its components are $(\omega \wedge \sigma)_{ab} = \omega_a \sigma_b - \omega_b \sigma_a = -(\omega \wedge \sigma)_{ba}$ and the wedge product itself has the property $\omega \wedge \sigma = -\sigma \wedge \omega$. Any 2-form can now be expressed as

$$\omega = \frac{1}{2!}\omega_{ab} \mathrm{d}x^a \wedge \mathrm{d}x^b \tag{3.72}$$

because then

$$\omega(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2!}\omega_{ab}\left(U^a V^b - U^b V^a\right) = \omega_{ab} U^a V^b. \tag{3.73}$$

This idea can be extended to p-forms in a natural way. A 3-form will be expressed in terms of a totally antisymmetric set of components ω_{abc} as

$$\omega = \frac{1}{3!}\omega_{abc} \mathrm{d}x^a \wedge \mathrm{d}x^b \wedge \mathrm{d}x^c \tag{3.74}$$

where the multiple wedge product is given by

$$\mathrm{d}x^a \wedge \mathrm{d}x^b \wedge \mathrm{d}x^c = \mathrm{d}x^a \otimes \mathrm{d}x^b \otimes \mathrm{d}x^c - \mathrm{d}x^b \otimes \mathrm{d}x^a \otimes \mathrm{d}x^c + \cdots . \tag{3.75}$$

The right-hand side is a sum of $3! = 6$ terms, giving all the permutations of (a, b, c), with a $+$ sign for each even permutation and a $-$ sign for each odd permutation, and the extension to higher p should be obvious. By adopting the rule that

$$(\mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_p}) \wedge (\mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_q}) = \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_p} \wedge \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_q} \tag{3.76}$$

we arrive at a definition of the wedge product, or *exterior product*, of a p-form ω and a q-form σ

$$\omega \wedge \sigma = \frac{1}{p!q!}\omega_{a_1\ldots a_p}\sigma_{b_1\ldots b_q}\mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_p} \wedge \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_q}. \tag{3.77}$$

The coordinate-free version of this definition is that, presented with the sequence of vector arguments $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{p+q})$, the $(p+q)$-form $\omega \wedge \sigma$ has the value

$$\omega \wedge \sigma(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{p+q}) = \frac{1}{p!q!}\sum_{\mathrm{P}} S(\mathrm{P})\, \omega\left(\boldsymbol{V}_{\mathrm{P}(1)}, \ldots, \boldsymbol{V}_{\mathrm{P}(p)}\right) \sigma\left(\boldsymbol{V}_{\mathrm{P}(p+1)}, \ldots, \boldsymbol{V}_{\mathrm{P}(p+q)}\right). \tag{3.78}$$

The news that I do not plan to wield this expression in anger may be greeted by some readers with relief, but it is not as bad as it looks. The labels $1, \ldots, (p+q)$ label a sequence of vectors, not their components, and the set $\{\mathrm{P}(1), \ldots, \mathrm{P}(p+q)\}$ is a permutation of these labels. The sum is over all these permutations P, and $S(\mathrm{P})$ is equal to 1 if P is an even permutation and -1 if P is an odd permutation. It should be quite straightforward to show that the exterior product is associative, $(\omega \wedge \sigma) \wedge \xi = \omega \wedge (\sigma \wedge \xi)$, and that, if ω is a p-form and σ a q-form, then $\omega \wedge \sigma = (-1)^{pq}\sigma \wedge \omega$.

A simple example of this machinery is afforded by the 'cross product' $\boldsymbol{u} \times \boldsymbol{v}$ of two vectors which, in elementary 3-dimensional vector algebra (using Cartesian coordinates) is defined to have the components

$$\boldsymbol{u} \times \boldsymbol{v} = \left((u^2v^3 - u^3v^2), (u^3v^1 - u^1v^3), (u^1v^2 - u^2v^1)\right). \tag{3.79}$$

It is easily seen that the three independent 2-forms $\mathrm{d}x^2 \wedge \mathrm{d}x^3$, $\mathrm{d}x^3 \wedge \mathrm{d}x^1$and $\mathrm{d}x^1 \wedge \mathrm{d}x^2$ with the arguments $(\boldsymbol{u}, \boldsymbol{v})$ produce exactly these components, but not in the form of a vector. We can combine them into the components of a *one-form*, by using the 3-dimensional Levi-Civita symbol

$$(\boldsymbol{u} \times \boldsymbol{v})_a = \tfrac{1}{2}\epsilon_{abc}\mathrm{d}x^b \wedge \mathrm{d}x^c(\boldsymbol{u}, \boldsymbol{v}) \tag{3.80}$$

and then, if we wish, use the Euclidean metric to convert this into a vector:

$$(\boldsymbol{u}\times\boldsymbol{v})^a = \tfrac{1}{2}g^{ab}\epsilon_{bcd}\mathrm{d}x^c\wedge\mathrm{d}x^d(\boldsymbol{u},\boldsymbol{v}). \tag{3.81}$$

There may seem to be a puzzle here. According to our definition, a one-form takes a vector argument to produce a scalar value, yet here the values $\mathrm{d}x^a(\boldsymbol{U}) = U^a$ seem to be the components of a vector. Indeed, according to the discussion following (3.64), these values *must* transform as the components of a vector. How can this be? Consider an observer, Olivia, who measures the component v^1 of the velocity of a particle relative to her own frame of reference. Her apparatus, which takes the velocity vector $\boldsymbol{v}$ and returns the number v^1, is a physical manifestation of the one-form $\mathrm{d}x^1$. But is this value a component of a vector, or is it a scalar? Other observers (say, Oliver and Orson) have their own frames of reference, with $x^{1'}$ and $x^{1''}$ axes that point in different directions. Their values, $v^{1'}$ and $v^{1''}$, are related to v^1 by the familiar coordinate transformations, and in this sense v^1, $v^{1'}$ and $v^{1''}$ are components of the same vector relative to different coordinate systems. On the other hand, the quantity that we can call 'Olivia's result for v^1' is a single number, whose value can be agreed on by all. In this sense it is a legitimate scalar. We see that, although the value of $\mathrm{d}x^1$ is a scalar, the definition of $\mathrm{d}x^1$ is tied to a particular coordinate system. If we regard $\mathrm{d}x^1$ as a fixed one-form then it has a fixed, scalar value when presented with a given vector. However, if we compare the value produced by $\mathrm{d}x^1$ with those that would be produced, given the same vector, by other one-forms, $\mathrm{d}x^{1'}$, $\mathrm{d}x^{1''}$ defined in an analogous way, but with respect to other coordinate systems, then these different scalar quantities will be related in the same way as the components of a vector, referred to the various coordinate systems.

The example of the cross product has two features that are worth elaborating on. In one sense, it is an object unique to 3-dimensional geometry, for the following reason. The components of a p-form, $\omega_{a_1\ldots a_p}$ are totally antisymmetric. How many independent components are there? Well, the p indices $a_1,\ldots,a_p$ must all have different values, and in d dimensions there are d values to choose from. The number of possible choices is the binomial coefficient $\binom{d}{p} = d!/p!(d-p)!$, so this is the number of independent components, and also the number of independent basis p-forms $\mathrm{d}x^{a_1}\wedge\cdots\wedge\mathrm{d}x^{a_p}$. Obviously, we get the same number of independent components for a $(d-p)$-form. Now, we obtained the cross product by presenting the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ as arguments to the basis 2-forms $\mathrm{d}x^a\wedge\mathrm{d}x^b$, of which there are $\binom{3}{2} = 3$. In d dimensions, the same procedure would lead to a set of $\binom{d}{2}$ components, and these can be assembled into a vector or a 1-form only if $\binom{d}{2} = d$, which is true only for $d = 3$. Thus, the notion of a cross product of two vectors that is itself a vector applies only in three dimensions. If, however, we do *not* insist that the resulting object be a vector, then an interesting and useful generalization is possible.

This brings me to the second feature of the cross product, which is that it illustrates the general notion of *dual tensors*. It would clearly be natural to regard the objects $U^a V^b - U^b V^a$ as the components not of a vector or a 1-form, but rather of an antisymmetric rank $\binom{2}{0}$ tensor. By analogy with a p-form, a totally antisymmetric rank $\binom{p}{0}$ tensor may be called a p-vector, but we must be careful not to confuse this terminology with the quite different notion of a 4-vector in special relativity. The number of independent components of a p-vector is the same as for a p-form and, while these components *might* be constructed, as in the cross product, from those of p vectors, they need not be. Given a p-vector, $V^{a_1 \dots a_p}$, we can generalize the second stage of our construction of the cross product by using the Levi-Civita symbol to create the $(d-p)$-form $^*\boldsymbol{V}$, which has components

$$^*V_{a_1 \dots a_{d-p}} = \frac{\hat{\omega}}{p!} \epsilon_{a_1 \dots a_{d-p} b_1 \dots b_p} V^{b_1 \dots b_p}. \tag{3.82}$$

The extra factor that I have called $\hat{\omega}$ here is needed to make sure that $^*\boldsymbol{V}$ is a genuine tensor. As explained in appendix A, the Levi-Civita symbol transforms as a tensor density of weight 1, with an extra factor of $\det|\Lambda|$, and the transformation of $\hat{\omega}$ must cancel this factor. In a manifold equipped with a metric, a natural choice is $\hat{\omega} = \sqrt{|g|}$, where $g = \det(g_{ab})$, so if we restrict ourselves to Cartesian coordinates in Euclidean space or Minkowski spacetime, then $\hat{\omega} = 1$. If we want to define dual tensors in a manifold without a metric, then we can do so by choosing a d-form, with components $\omega_{a_1 \dots a_d} = \omega_{1 \dots d} \epsilon_{a_1 \dots a_d}$, and setting $\hat{\omega}$ equal to its one independent component $\omega_{1 \dots d}$. The meaning of 'duality' will then depend on which d-form we have chosen to play this special role. Note that, since both the p-vector $\boldsymbol{V}$ and the $(d-p)$-form $^*\boldsymbol{V}$ have $\binom{d}{p}$ independent components, there is exactly enough information in $\boldsymbol{V}$ to construct $^*\boldsymbol{V}$ and *vice versa*. That being so, we might expect that the process can be reversed to convert a p-form ω into a $(d-p)$-vector $^*\omega$. Indeed it can, and the components of $^*\omega$ are

$$^*\omega^{a_1 \dots a_{d-p}} = \frac{\hat{\omega}^{-1}}{p!} \epsilon^{a_1 \dots a_{d-p} b_1 \dots b_p} \omega_{b_1 \dots b_p}. \tag{3.83}$$

Equally, we might guess that the tensor dual to $^*\boldsymbol{V}$ is $\boldsymbol{V}$. The correct relation turns out to be $^{**}\boldsymbol{V} = (-1)^{p(d-p)}\boldsymbol{V}$, and similarly $^{**}\omega = (-1)^{p(d-p)}\omega$ (see exercise 3.7). The duality operation represented by * is called the 'Hodge star' operation.

An important example of a 2-form in Minkowski spacetime is the electromagnetic field strength tensor (3.50). I shall show shortly that Maxwell's equations can be expressed in a compact and elegant form by using this tensor and its dual, but to do this, we need a further new idea. The *exterior derivative* d is a differential operator, which is nicely illustrated by the way in which the 2-form F, whose components are $F_{\mu\nu}$, is obtained from the 1-form vector potential A. The operator d is defined so as to produce from a p-form ω a $(p+1)$-form $\mathrm{d}\omega$. For this purpose, it is convenient to regard a scalar field f as a 0-form, in which case $\mathrm{d}f$ is the gradient that we have already met:

$$\mathrm{d}f = \frac{\partial f}{\partial x^a}\mathrm{d}x^a. \tag{3.84}$$

The notation here is quite consistent. If we take a special scalar field which, in a suitable coordinate system, can be expressed as $f(x) = x^1$, say, then $\mathrm{d}f = (\partial x^1/\partial x^a)\mathrm{d}x^a = \delta^1_a \mathrm{d}x^a = \mathrm{d}x^1$. The action of d on a 1-form $\omega = \omega_a \mathrm{d}x^a$ is

$$\mathrm{d}\omega = \frac{\partial \omega_a}{\partial x^b}\mathrm{d}x^b \wedge \mathrm{d}x^a = -\omega_{a,b}\mathrm{d}x^a \wedge \mathrm{d}x^b \tag{3.85}$$

where I have used the antisymmetry of the wedge product and the comma notation from chapter 2 for partial derivatives. Now, a 2-form is supposed to have antisymmetric components, $(\mathrm{d}\omega)_{ab} = -(\mathrm{d}\omega)_{ba}$. In general, $\omega_{a,b}$ will not be equal to $-\omega_{b,a}$, but because of the antisymmetry of $\mathrm{d}x^a \wedge \mathrm{d}x^b$, only the antisymmetric combination $\omega_{b,a} - \omega_{a,b}$ actually contributes to $\mathrm{d}\omega$. Since a and b are dummy summation variables in (3.85), we can rename them as b and a to get

$$\mathrm{d}\omega = -\omega_{b,a}\mathrm{d}x^b \wedge \mathrm{d}x^a = +\omega_{b,a}\mathrm{d}x^a \wedge \mathrm{d}x^b \tag{3.86}$$

and therefore

$$\mathrm{d}\omega = \tfrac{1}{2}(\omega_{b,a} - \omega_{a,b})\mathrm{d}x^a \wedge \mathrm{d}x^b. \tag{3.87}$$

In view of the general expression (3.72) the components of $\mathrm{d}\omega$ are actually the antisymmetric quantities $(\mathrm{d}\omega)_{ab} = \omega_{b,a} - \omega_{a,b}$. Evidently, the electromagnetic field strength (3.50) can be written in coordinate-free language simply as $F = \mathrm{d}A$. (Readers should also have little difficulty in convincing themselves that in 3-dimensional Euclidean geometry the curl of a vector field $\boldsymbol{\nabla} \times \boldsymbol{v}$ can be constructed using d in much the same way as the cross product of two vectors.)

In general, the action of d on a p-form ω is

$$\mathrm{d}\omega = \frac{1}{p!}\left(\partial_b \omega_{a_1 \ldots a_p}\right)\mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_p} \tag{3.88}$$

and this could be rewritten in a totally antisymmetric form analogous to (3.87). Using the definition of the exterior product (3.77), it is not hard to show that d obeys a modified version of the Leibniz rule: for a p-form ω and a q-form σ,

$$\mathrm{d}(\omega \wedge \sigma) = \mathrm{d}\omega \wedge \sigma + (-1)^p \omega \wedge \mathrm{d}\sigma. \tag{3.89}$$

Consider, in particular, the case that ω is itself the exterior derivative of a $(p-1)$-form, say $\omega = \mathrm{d}\sigma$. Each component of $\mathrm{d}\omega$ will be a sum of terms of the form $(\partial_a \partial_b - \partial_b \partial_a)\sigma_{c\ldots}$, which are identically zero. Thus, for any p-form, we have $\mathrm{d}^2\omega = 0$. The mathematical jargon for this says that the operator d is *nilpotent*. In 3-dimensional Euclidean geometry, the well-known identities $\boldsymbol{\nabla} \times (\boldsymbol{\nabla}\phi) = 0$ and $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}) = 0$, valid for any scalar field ϕ and any vector field $\boldsymbol{v}$ can be understood in terms of the identity $\mathrm{d}^2 = 0$. As far as Maxwell's equations are concerned, the two equations (3.45) and (3.46) are equivalent to the statement

$$\mathrm{d}F = 0. \tag{3.90}$$

Usually, given that $\mathrm{d}^2 = 0$, we take this to imply that F can be expressed in terms of a vector potential as $F = \mathrm{d}A$, but there is a subtlety here. Suppose that a p-form ω satisfies $\mathrm{d}\omega = 0$. It is said to be *closed*. According to a theorem known as the *Poincaré lemma*, we can always find a $(p-1)$-form σ such that $\omega = \mathrm{d}\sigma$, *provided* that we restrict attention to a sufficiently simple region of the manifold on which ω is defined; an open set that is topologically equivalent to the interior of the unit sphere in $\mathbb{R}^d$ will do. If ω can be expressed as $\mathrm{d}\sigma$, then it is said to be *exact*, so the Poincaré lemma says that any closed form is 'locally exact'. However, a closed form may not be exact over the whole manifold. That is to say, although we can express ω as $\mathrm{d}\sigma$ in any local region of the appropriate kind, there may not be a single σ that works throughout the whole manifold. This depends on the global topology of the manifold, and one way of characterizing this global topology is in terms of those forms that are closed, but not exact. Roughly speaking, this constitutes what is called the *cohomology* of the operator d. In electromagnetism, the Maxwell equation (3.45) forbids the existence of magnetic monopoles *unless* we allow for the possibility that a single 1-form potential A may not be valid through the whole of spacetime, and I shall take up this question again in chapter 13.

To express the remaining Maxwell equations (3.44) and (3.47) in our new language, we start from the contravariant version of the field strength tensor (3.51) which, according to our present terminology is a 2-vector $\boldsymbol{F}$. Its dual is a 2-form ${}^*\boldsymbol{F}$, whose components are

$${}^*F_{\mu\nu} = \begin{pmatrix} 0 & -B^1 & -B^2 & -B^3 \\ B^1 & 0 & -E^3 & E^2 \\ B^2 & E^3 & 0 & -E^1 \\ B^3 & -E^2 & E^1 & 0 \end{pmatrix}. \tag{3.91}$$

Notice that duality has the effect of interchanging electric and magnetic fields, and that this would be a symmetry of Maxwell's equations in the absence of charged particles. The exterior derivative $\mathrm{d}^*\boldsymbol{F}$ is a 3-form, whose components are

$$(\mathrm{d}^*\boldsymbol{F})_{\mu\nu\sigma} = \partial_\mu {}^*F_{\nu\sigma} + \partial_\nu {}^*F_{\sigma\mu} + \partial_\sigma {}^*F_{\mu\nu}. \tag{3.92}$$

It is a simple matter to check that these are totally antisymmetric, owing to the antisymmetry of ${}^*\boldsymbol{F}$. The electromagnetic current is a vector $\boldsymbol{j}$ and its dual tensor is a 3-form, with components ${}^*j_{\mu\nu\sigma} = \epsilon_{\mu\nu\sigma\tau} j^\tau$. Each of these 3-forms has, as we saw above, only $\binom{4}{3} = 4$ independent components; for example, ${}^*j_{012} = j^3$. Thus, the tensor equation

$$\mathrm{d}^*\boldsymbol{F} = c^{-1*}\boldsymbol{j} \tag{3.93}$$

is a set of four equations, which are equivalent to the Maxwell equations (3.44) and (3.47). For example,

$$(\mathrm{d}^*\boldsymbol{F})_{012} = \partial_0{}^*F_{12} + \partial_1{}^*F_{20} + \partial_2{}^*F_{01} = \left(\nabla \times \boldsymbol{B} - \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t}\right)^3 = \frac{1}{c}j^3. \quad (3.94)$$

While Maxwell's equations as expressed by (3.90) and (3.93) are somewhat more compact than the original versions, readers may well feel that this is more than offset by the amount of space needed to say what the notation means! However, the compactness of the notation for dealing with antisymmetric tensors and the fact that these equations are now in a completely coordinate-free form bring significant advantages when one is dealing, for example, with the non-Abelian generalizations of electromagnetism that I shall discuss in chapters 8 and 12 or with manifolds that are more complicated than Minkowski spacetime.

3.7.3 Configuration space and its relatives

By now, it should come as no surprise that the antisymmetric structure of the Poisson bracket (3.18) has a geometrical interpretation in terms of differential forms. The version of this interpretation that I plan to explain applies to non-relativistic physics, in which physical events are regarded as taking place in a 3-dimensional space, rather than in a 4-dimensional spacetime. Relativistic versions are possible, but they involve subtleties in which I do not want to get embroiled. For present purposes, then, we regard time not as a coordinate but as a parameter that labels points on the path of a particle through *space*. For a system of N particles, it becomes a little awkward to deal with N paths, all labelled by the same parameter. It is more convenient to deal instead with a $3N$-dimensional manifold, in which a single point represents the positions of all the particles. The $3N$ generalized coordinates $\{q^i\}$ introduced in §3.1 serve as coordinates on this manifold, which is called *configuration space*, and which I will denote by $\mathbb{Q}$. A possible history of the entire system corresponds to a single path through this manifold. However, a point in configuration space does not represent a unique state of the system. To do that, we have to take account either of the velocities of the particles or of their momenta as well as their positions.

From a geometrical point of view, the natural way of doing this is to construct a suitable manifold, which is an example of a *fibre bundle* analogous, but by no means identical, to the Galilean spacetime illustrated in figure 2.13. Consider first how we might take account of velocities. Given a point P in configuration space and a curve passing through it that represents a possible history of the system, the $3N$ generalized velocities $\{\dot{q}^i\}$ that the particles have at the instant when their positions correspond to P are the components of the tangent vector $\mathrm{d}/\mathrm{d}t$ to this curve at P. The set of all tangent vectors at P (or, equivalently, the set of tangent vectors to all possible curves through P) forms a *vector space*, called the *tangent space* to $\mathbb{Q}$ at the point P and denoted by $T_P\mathbb{Q}$. (The precise mathematical definition of a vector space is given in appendix A, but for the

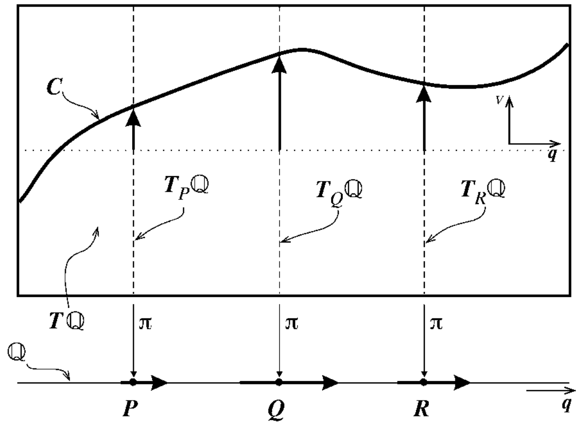


Figure 3.1. A one-dimensional configuration space $\mathbb{Q}$, with coordinate q, and its tangent bundle $T^*\mathbb{Q}$, with coordinates q and v.

purposes of our present discussion, readers' intuition gained from the elementary study of Euclidean vectors should serve just as well.)

We now construct a new manifold, called the *tangent bundle* of $\mathbb{Q}$ and denoted by $T\mathbb{Q}$. Intuitively, we can think of doing this by 'bundling up' the tangent spaces at all points of $\mathbb{Q}$ to form a single object. This is illustrated in figure 3.1 for the only case that can easily be drawn, namely a single particle in one dimension, for which $\mathbb{Q}$ is just the real line. To be mathematically precise, we have do things the other way round, because we want $T\mathbb{Q}$ to be a differentiable manifold in its own right. Thus we say that $T\mathbb{Q}$ is a $6N$-dimensional manifold (though in figure 3.1 it has only two dimensions), topologically equivalent to $\mathbb{R}^{6N}$ and equipped with a *projection* π. This projection is a map which, for each point P of the configuration space $\mathbb{Q}$ picks out the $3N$-dimensional slice (or *fibre*) of $T\mathbb{Q}$ corresponding to $T_P\mathbb{Q}$ and maps each point of this slice to the appropriate point P of $\mathbb{Q}$. Given the existence of this projection, there is a natural way of setting up coordinates on the tangent bundle. That is, half the coordinates, $\{q^i\}$, serve to identify a slice of the bundle, corresponding to a point P in $\mathbb{Q}$ whose coordinates are $\{q^i\}$, while the other half, say $\{v^i\}$ identify a point within this slice corresponding to a possible set of velocities for the particles whose positions correspond to P. I will use $\{v^i\}$ to denote these coordinates in the tangent bundle and $\{\dot{q}^i(t)\}$ for the actual velocities corresponding to a specific state of the system of particles. In figure 3.1, I found it impossible to draw a 1-dimensional curve inside the 1-dimensional configuration space $\mathbb{Q}$, but the arrows at P, Q and R represent the tangent vectors to such a curve at these points. The vector field in $\mathbb{Q}$ that comprises all these tangent vectors gives rise to a curve C in the tangent bundle which, for reasons that should be apparent, is called a *cross section* of the bundle. Each point on C now represents a unique state of the system, being specified by both positions and velocities. (In higher dimensions, a vector field on $\mathbb{Q}$ would correspond to a family of curves representing a family of possible histories of the system, but I will not develop this point in detail.)

The Lagrangian $L(\{q^i\}, \{v^i\})$ is a scalar field defined on the tangent bundle $T\mathbb{Q}$. It must be a genuine scalar, because it has a definite value for each state of the system and, therefore, at each point of $T\mathbb{Q}$, regardless of how we choose the generalized coordinates and velocities. To avoid tiresome complications, I shall deal with the most usual case in which L can be expressed as

$$L = \tfrac{1}{2} g_{ij}(q) v^i v^j - V(q). \tag{3.95}$$

The objects $g_{ij}(q)$ are the components of a metric tensor field *on the configuration space* $\mathbb{Q}$. This metric is, of course, related to that of the ordinary 3-dimensional space from which we started. For example, if we consider two particles in Euclidean space, whose positions in Cartesian coordinates are $\boldsymbol{x}$ and $\boldsymbol{y}$, and whose masses are m_1 and m_2, then we can choose generalized coordinates $(q^1, \ldots, q^6) = (x^1, x^2, x^3, y^1, y^2, y^3)$, in which case g_{ij} is diagonal, with elements $(m_1, m_1, m_1, m_2, m_2, m_2)$, related in an obvious way to the Euclidean metric δ_{ab} $(a, b = 1, \ldots, 3)$. More generally, g_{ij} may depend on the positions q, either because we want to think about a non-Euclidean space or because we are not using Cartesian coordinates. The generalized momenta p_i conjugate to q^i are

$$p_i = \frac{\partial L}{\partial v^i} = g_{ij}(q) v^j. \tag{3.96}$$

We see that they are obtained by lowering the indices of the components v^j of a vector field on the configuration space $\mathbb{Q}$, and are therefore themselves the components of a 1-form field, or of a 1-form if we restrict our attention to a particular point P. Now, the set of all 1-forms at P forms a vector space, called the *cotangent space* $T_P^*\mathbb{Q}$, and we can bundle together all the cotangent spaces at different points to form the *cotangent bundle* $T^*\mathbb{Q}$ just as we previously constructed the tangent bundle. On this manifold, a natural set of coordinates is provided by the $6N$ quantities $(\{q^i\}, \{p_i\})$.

3.7.4 The symplectic geometry of phase space

The fibre bundle $T^*\mathbb{Q}$ is known to physicists as *phase space*. Since it is a differentiable manifold, we might well choose to place on it a system of coordinates ξ^α, the index α running from 1 to $6N$, with associated bases $\partial/\partial\xi^\alpha$ and $\mathrm{d}\xi^\alpha$ for vector and 1-form fields. For the most part, it will prove sensible to retain the natural division of these coordinates into q^i and p_i, with i running from 1 to $3N$. The lower indices on the p_i are inherited from the role of these quantities as the components of a 1-form field *on configuration space* $\mathbb{Q}$ rather than as coordinates on phase space. It is worth observing, though, that a change of coordinates in $\mathbb{Q}$ with, say, $\Lambda^{i'}{}_i = \partial q^{i'}/\partial q^i$ leads to a corresponding change of coordinates in phase space, in which the momenta still transform 'covariantly' as $p_{i'} = \Lambda^{i}{}_{i'} p_i$. This means that the large transformation matrix $\mathbf{\Lambda}^{\alpha'}{}_{\alpha} = \partial\xi^{\alpha'}/\partial\xi^{\alpha}$ is constructed from both of the matrices $\Lambda^{i'}{}_{i}$ and $\Lambda^{i}{}_{i'}$, and readers may enjoy finding out for themselves exactly how this works.

Normally, a manifold has a useful application in physics only when we endow it with some geometrical structure that is apposite for the phenomena we want to describe, so the central question that now arises is, what is the natural geometrical structure for phase space? In principle, we might try to endow phase space with a metric, but this is unlikely to be of much use because an expression such as $(\Delta q)^2 + (\Delta p)^2$ has, except by accident, no sensible meaning. The structure that turns out to be meaningful in the context of Hamiltonian dynamics is one that we have not yet met. It is called a *symplectic* structure. In the same way that the metrical structure of a relativistic spacetime is implemented by a special rank $\binom{0}{2}$ tensor g, the symplectic structure of phase space is implemented by a rank $\binom{0}{2}$ tensor Ω. The difference is that while g is symmetric, Ω is antisymmetric: it is called the *symplectic 2-form.* With our preferred system of coordinates, it is

$$\Omega = \mathrm{d}q^i \wedge \mathrm{d}p_i, \tag{3.97}$$

where a sum over $i = 1, \ldots, 3N$ is implied, as the notation suggests. The meanings of the 1-forms $\mathrm{d}q^i$ and $\mathrm{d}p_i$ are the same as in (3.64), but we have now split our coordinates into two sets. Thus, a vector field on phase space will have 'q-type' and 'p-type' components, say

$$\boldsymbol{V} = V^i \frac{\partial}{\partial q^i} + \tilde{V}_i \frac{\partial}{\partial p_i} \tag{3.98}$$

and the values of the basis 1-forms when presented with this vector field are

$$\mathrm{d}q^i(\boldsymbol{V}) = V^i \qquad \text{and} \qquad \mathrm{d}p_i(\boldsymbol{V}) = \tilde{V}_i. \tag{3.99}$$

The 2-form Ω is actually the exterior derivative of what is called the *canonical* 1*-form* $\theta = p_i \mathrm{d}q^i$. In fact, the rule (3.85) tells us that $\Omega = -\mathrm{d}\theta$. The significance of this is the following. Given a curve with tangent vector $\mathrm{d}/\mathrm{d}t$ that represents a history of our system, the velocities are $\dot{q}^i(t) = \mathrm{d}q^i(\mathrm{d}/\mathrm{d}t)$. Thus, the scalar quantity $\theta(\mathrm{d}/\mathrm{d}t) = p_i\dot{q}^i$ is what appears in the Legendre transformation (3.14) that enables us to move from a Lagrangian to a Hamiltonian description of the system. Its geometrical manifestation θ plays the analogous role when we move from a description in terms of the tangent bundle to one in terms of the cotangent bundle.

Like the metric tensor, the symplectic 2-form can be used to define a correspondence between vectors and 1-forms on phase space. Given a vector $\boldsymbol{V}$, the object $\omega_{\boldsymbol{V}} = \Omega(\boldsymbol{V}, \quad)$ is a 1-form, because it can accept one more vector argument to produce a scalar. In components, the $6N$ quantities $(\omega_{\boldsymbol{V}})_\beta = \Omega_{\alpha\beta}V^\alpha$ are the components of a unique 1-form associated with the vector $\boldsymbol{V}$. Can we invert this to find a unique vector $\boldsymbol{V}_\omega$ associated with a given one-form ω? In other words, do the equations $\omega_\beta = \Omega_{\alpha\beta}(\boldsymbol{V}_\omega)^\alpha$ have a unique solution for the components of $\boldsymbol{V}_\omega$? The answer is yes, provided that the matrix $\Omega_{\alpha\beta}$ has an inverse, which means that its determinant is nonzero. If this condition is met, then Ω is said to be *non-degenerate*. Mathematically, this property is normally insisted on as part of the definition of a symplectic structure. In the case we have considered, the 2-form defined by (3.97) is indeed non-degenerate. If we arrange our coordinates in the order $(\xi^1, \dots, \xi^{6N}) = (q^1, \dots, q^{3N}, p_1, \dots, p_{3N})$, then the components of Ω are

$$\Omega_{\alpha\beta} = \begin{pmatrix} 0 & \mathbb{I} \\ -\mathbb{I} & 0 \end{pmatrix} \tag{3.100}$$

where $\mathbb{I}$ is the $3N \times 3N$ unit matrix. Each row and each column of this matrix has exactly one nonzero element and its determinant is either 1 or -1. There are, however, important physical examples in which Ω is degenerate. This typically indicates a mismatch between the numbers of coordinates and momenta and comes about when there are ‘unphysical’ degrees of freedom, such as the gauge degrees of freedom in electromagnetism. A Hamiltonian description of the dynamics of such systems is often possible, but requires special techniques that are beyond the scope of the present discussion. (A simple example is discussed by Lawrie and Epp (1996).)

The application of the general idea of symplectic geometry to Hamiltonian dynamics depends on our identifying a special class of vector fields on phase space, namely those whose associated 1-forms $\Omega(\boldsymbol{V},\ \ )$ are the gradients of scalar quantities that represent physical properties of our system. That is to say, given a quantity $A(\{q^i\}, \{p_i\})$, we can associate with it a vector field $\boldsymbol{V}_A$ such that

$$\Omega(\boldsymbol{V}_A,\ \ ) = \mathrm{d}A. \tag{3.101}$$

A vector field for which this equation can be solved to find the corresponding scalar A is called a *Hamiltonian vector field*, although A is not necessarily the Hamiltonian. Let us find the components of $\boldsymbol{V}_A$. Using the definitions of the wedge product and the exterior derivative, we can write (3.101) in components as

$$V_A^i \mathrm{d}p_i - \tilde{V}_i^A \mathrm{d}q^i = \frac{\partial A}{\partial q^i}\mathrm{d}q^i + \frac{\partial A}{\partial p_i}\mathrm{d}p_i. \tag{3.102}$$

We see that $V_A^i = \partial A/\partial p_i$ and $\tilde{V}_i^A = -\partial A/\partial q^i$, and so

$$\boldsymbol{V}_A = \frac{\partial A}{\partial p_i}\frac{\partial}{\partial q^i} - \frac{\partial A}{\partial q^i}\frac{\partial}{\partial p_i}. \tag{3.103}$$

This is none other than the differential operator $-\{A,\ \ \}_\mathrm{P}$, of which we encountered examples in §3.4. The Poisson bracket itself is

$$\{A, B\}_\mathrm{P} = -\Omega(\boldsymbol{V}_A, \boldsymbol{V}_B) = \frac{\partial A}{\partial q^i}\frac{\partial B}{\partial p_i} - \frac{\partial B}{\partial q^i}\frac{\partial A}{\partial p_i}. \tag{3.104}$$

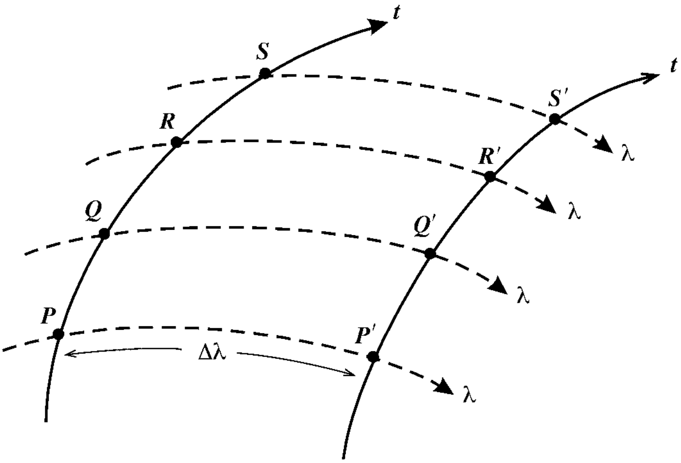


Figure 3.2. The curve $PQRS$ represents a possible trajectory of a system in phase space. The curves PP', QQ', RR' and SS' are integral curves of the Hamiltonian vector field $\boldsymbol{V}_A$ associated with a dynamical quantity $A(\{q^i\}, \{p_i\})$. If A is the conserved quantity corresponding to a symmetry of the system, then $P'Q'R'S'$ is also a possible trajectory.

Let us finally see how the time evolution and the symmetries of a Hamiltonian system appear from a geometrical point of view. Given a vector field $\boldsymbol{V}$, each point of phase space lies on exactly one of a family of curves, to which $\boldsymbol{V}$ gives the tangent vectors. They are called the *integral curves* of $\boldsymbol{V}$. The physical constitution of a system (the forces that act between its particles, and so on) is specified by selecting a function $H(\{q^i\}, \{p_i\})$ as the Hamiltonian and by identifying the parameter t that labels points on the integral curves of $\boldsymbol{V}_H$ as time. Thus we have

$$\boldsymbol{V}_H = \frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i}\frac{\partial}{\partial p_i} \tag{3.105}$$

which reproduces the equation of motion (3.17). We see that the integral curves of $\boldsymbol{V}_H$ are the possible trajectories through phase space of the point that represents the state of the system as it evolves with time.

To appreciate the role of symmetries, we need to know the commutator $[\boldsymbol{V}_A, \boldsymbol{V}_B]$ of two vector fields, regarded simply as differential operators. A few lines of algebra suffice to verify that

$$[\boldsymbol{V}_A, \boldsymbol{V}_B] = -\boldsymbol{V}_C \tag{3.106}$$

where $C = \{A, B\}_\mathrm{P}$. Thus, if $\{A, B\}_\mathrm{P} = 0$, then $[\boldsymbol{V}_A, \boldsymbol{V}_B] = 0$ and the two vector fields commute. Now look at figure 3.2. The solid curve passing through the points P, Q, R and S is an integral curve of $\boldsymbol{V}_H$—a possible trajectory of the system through phase space. The dashed curves are the integral curves of the vector field $\boldsymbol{V}_A = \mathrm{d}/\mathrm{d}\lambda$, associated with some physical quantity A, that pass through $P, \ldots, S$. The points $P', \ldots, S'$ are found by displacing $P, \ldots, S$ by the same parameter distance $\Delta\lambda$ along the dashed curves. This corresponds to a translation of the system of the kind that we studied in earlier sections. For example, let $\boldsymbol{n}$ be a unit vector in ordinary 3-dimensional space and $\boldsymbol{p}$ the momentum of a particle. Then $\boldsymbol{n} \cdot \boldsymbol{p}$ is the component of momentum in the direction of $\boldsymbol{n}$. In the $3N$-dimensional configuration space for N particles, there is a vector with $3N$ components n^i, consisting of N copies of $\boldsymbol{n}$ and the quantity $A = n^i p_i = \sum_{j=1}^{N} \boldsymbol{n} \cdot \boldsymbol{p}_j$, where j labels the N particles, is the component of the total momentum in the direction of $\boldsymbol{n}$. The corresponding Hamiltonian vector field is $\boldsymbol{V}_A = n^i \partial/\partial q^i$ and the displacement corresponds to a space translation of the whole system by a distance $\Delta\lambda$ in the direction specified by $\boldsymbol{n}$. If $[\boldsymbol{V}_A, \boldsymbol{V}_H] = 0$, then the curve passing through the displaced points $P', \ldots, S'$ will be another integral curve of $\boldsymbol{V}_H$—another possible trajectory of the system through phase space. (I shall not prove this assertion. Enterprising readers may like to attempt a proof, or to consult, for example, Schutz (1980) where the relevant concept of a *Lie derivative* is explained in detail.) This is a special situation: while there is certainly a trajectory passing through P', this trajectory need not, in general, pass through $Q', \ldots, S'$. When it does, we can conclude that the system has a symmetry: the Hamiltonian is unchanged by the displacement and the displaced system evolves with time in the same way as the original one. The condition $[\boldsymbol{V}_A, \boldsymbol{V}_H] = 0$ that makes this true is equivalent to $\{A, H\}_{\mathrm{P}} = 0$ and this, as we know, means that A is a conserved quantity. But we can now appreciate this result in a slightly different light, because the conditions are the same if we interchange A and H. In terms of figure 3.2, we can say that if H is unchanged by a displacement along the integral curves of $\boldsymbol{V}_A$ (so H has a symmetry) then, by the same token, A is unchanged by a displacement along the integral curves of $\boldsymbol{V}_H$ (so A is constant in time). In fact, we can say more. Since $\mathrm{d}A/\mathrm{d}\lambda = \{A, A\}_{\mathrm{P}} = 0$, the quantity A is constant along the integral curves of $\boldsymbol{V}_A$ as well. Thus, the integral curves of $\boldsymbol{V}_A$ and $\boldsymbol{V}_H$ mesh together to form surfaces in phase space, and both A and H are constant over any one of these surfaces. This is an example of a more general result known as *Frobenius' theorem*, which is also discussed by Schutz (1980).

Exercises

3.1. Express the Lagrangian $L = \frac{1}{2}m\dot{\boldsymbol{x}}^2 - V(\boldsymbol{x})$ for a single particle in cylindrical coordinates (r, θ, z) with $x = r\cos\theta$ and $y = r\sin\theta$. Show that the generalized momentum conjugate to θ is the angular momentum $mr^2\dot{\theta}$ about the z axis. If the potential V has cylindrical symmetry (that is, it is independent of θ), show, by considering the transformation $\theta \to \theta + \epsilon$, that the conserved quantity F in (3.12) is the angular momentum. When ϵ is infinitesimal, find the corresponding transformation of the Cartesian coordinates x and y. Working in Cartesian coordinates, show that if the Lagrangian is invariant under this transformation, then the conserved quantity is the z component of the angular momentum $\boldsymbol{J} = \boldsymbol{x} \times \boldsymbol{p}$. Show that if the potential is spherically symmetric (that is, it is a function only of $x^2 + y^2 + z^2$), then all three components of angular momentum are conserved. In cylindrical coordinates, show that the generator of rotations about the z axis is $-\mathrm{i}\partial/\partial\theta$. In Cartesian coordinates, show that the rotation generators are $\boldsymbol{\mathcal{J}} = \mathrm{i}\{\boldsymbol{J},\ \}_{\mathrm{P}} = \boldsymbol{x} \times \boldsymbol{\mathcal{P}}$.

3.2. Consider the Lagrangian $L = \frac{1}{2}m\dot{\boldsymbol{x}}^2 - V(\boldsymbol{x})$ and the Hamiltonian $H = (1/2m)\boldsymbol{p}^2 + V(\boldsymbol{x})$. Show that Hamilton's equations are equivalent to the Euler–Lagrange equations together with the definition of the canonical momentum. Now consider the Lagrangian $L = \boldsymbol{p} \cdot \dot{\boldsymbol{x}} - (1/2m)\boldsymbol{p}^2 - V(\boldsymbol{x})$, where $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$ and $\boldsymbol{p}$ are to be treated as independent variables. Show that the Euler–Lagrange equations reproduce the previous equations of motion, together with the relation $\boldsymbol{p} = m\dot{\boldsymbol{x}}$.

3.3. For a single particle in Minkowski spacetime, show (taking careful account of the minus sign in (3.33)) that the Hamiltonian $H = -\eta_{\mu\nu}p^\mu\dot{x}^\nu - L$ expressed as a function of the momenta leads to a set of Hamilton's equations which reproduce the correct equation of motion together with the definition (3.33) of the momenta, provided that derivatives with respect to proper time are used. Show that this Hamiltonian is a conserved quantity, but is not equal to the total energy of the particle.

3.4. Using elementary kinetic theory for a non-relativistic ideal gas in its rest frame, show that $\langle p^i(\mathrm{d}x^j/\mathrm{d}t)\rangle = (p/n)\delta^{ij}$, where p^i and $\mathrm{d}x^i/\mathrm{d}t$ are the Cartesian components of momentum and velocity, p and n are the pressure and number density and the average $\langle\cdots\rangle$ is taken over all the particles. Assume that the same is true for a relativistic gas if the spatial components of the momentum in (3.34) are used. For the relativistic gas in its rest frame, imagine dividing the volume it occupies into cells, each of which is small compared with the total volume but still contains many particles. Define the average of the stress tensor (3.42) for each cell as

$$\langle T^{\mu\nu}\rangle = \int_{\mathrm{cell}} \mathrm{d}^3x\, T^{\mu\nu}(x)/\text{Volume of cell.}$$

Show that this average has the form shown in (3.43). More generally, consider a fluid whose stress tensor field has this form at the point x when measured relative to the rest frame of the fluid element at x. Show that its stress tensor field in any frame of reference is

$$T^{\mu\nu} = c^{-2}(\rho + p)u^\mu u^\nu - pg^{\mu\nu}$$

where $u^\mu(x)$ is the 4-velocity of the fluid element at x and $\rho(x)$ and $p(x)$ are the energy density and pressure as measured in the rest frame of this element.

3.5. Consider the Lagrangian density

$$\mathcal{L} = \tfrac{1}{4}F^{\mu\nu}F_{\mu\nu} - \tfrac{1}{2}F^{\mu\nu}(\partial_\mu A_\nu - \partial_\nu A_\mu) - c^{-1}j_{\rm e}^\mu A_\mu .$$

Derive two Euler–Lagrange equations, treating $F^{\mu\nu}$ and A_μ as independent variables, and show that they reproduce (3.50) and (3.52).

3.6. In a particular frame of reference, define the Lagrangian for electromagnetic fields as $L = -\frac{1}{4}\int \mathrm{d}^3x\, F_{\mu\nu}F^{\mu\nu}$. Show that $L = \frac{1}{2}\int \mathrm{d}^3x(\boldsymbol{E}^2 - \boldsymbol{B}^2)$. Define the generalized momentum conjugate to $A_\mu(x)$ as $\Pi^\mu(x) = \delta L/\delta(\partial_0 A_\mu)$, where $\delta/\delta(\cdots)$ is the functional derivative discussed in appendix A. Show that $\Pi^i = E^i$ for $i = 1, 2, 3$ and $\Pi^0 = 0$. Now define the Hamiltonian $H = \int \mathrm{d}^3x\, \Pi^\mu \partial_0 A_\mu - L$. Using Gauss' law $\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0$ (which is one of the Euler–Lagrange equations in the absence of charged particles), show that H is the integral over all space of the energy density $\frac{1}{2}(\boldsymbol{E}^2 + \boldsymbol{B}^2)$.

3.7. For a p-vector $\boldsymbol{V}$, the following is an outline proof that ${}^{**}V^{12\ldots p} = (-1)^{p(d-p)}V^{12\ldots p}$. Convince yourself that each step is correct:

$$\begin{aligned}
{}^{**}V^{1\ldots p} &= \frac{1}{p!(d-p)!}\epsilon^{1\ldots p b_1\ldots b_{d-p}}\epsilon_{b_1\ldots b_{d-p}a_1\ldots a_p}V^{a_1\ldots a_p}\\
&= \frac{1}{p!}\epsilon^{1\ldots d}\epsilon_{(p+1)\ldots d a_1\ldots a_p}V^{a_1\ldots a_p}\\
&= \epsilon_{(p+1)\ldots d 1\ldots p}V^{1\ldots p}\\
&= (-1)^{p(d-p)}V^{1\ldots p}
\end{aligned}$$

Convince yourself that the same result holds for every component of $\boldsymbol{V}$ and for every component of a p-form ω.

3.8. Two particles move in one dimension. Their positions are x^1 and x^2, their momenta are p_1 and p_2 and the Hamiltonian is

$$H = \frac{1}{2m}(p_1^2 + p_2^2) + \frac{k}{2}\left(x^1 - x^2\right)^2 .$$

To avoid complications, assume that these particles can pass through each other, so configurations with $x^1 < x^2$ and $x^1 > x^2$ are both allowed.

(a) Find the Hamiltonian vector fields $\boldsymbol{V}_H$ and $\boldsymbol{V}_P$, where $P = p_1 + p_2$ is the total momentum, in terms of the phase-space coordinates x^i and p_i. Verify that $\{P, H\}_{\rm P} = 0$.

(b) Define a new set of phase-space coordinates (X, P, ρ, θ) by

$$\begin{aligned}
x^1 &= X + \tfrac{1}{2}\rho\cos\theta & \qquad p_1 &= \tfrac{1}{2}\left(P + \sqrt{2km}\,\rho\sin\theta\right)\\
x^2 &= X - \tfrac{1}{2}\rho\cos\theta & \qquad p_2 &= \tfrac{1}{2}\left(P - \sqrt{2km}\,\rho\sin\theta\right).
\end{aligned}$$

and show that the symplectic 2-form is

$$\Omega = \mathrm{d}x^i \wedge \mathrm{d}p_i = \mathrm{d}X \wedge \mathrm{d}P + \sqrt{km/2}\,\rho \mathrm{d}\rho \wedge \mathrm{d}\theta.$$

(c) Express H in terms of these coordinates and show that

$$\boldsymbol{V}_P = \frac{\partial}{\partial X} \qquad \boldsymbol{V}_H = \frac{1}{2m} P \frac{\partial}{\partial X} - \sqrt{\frac{2k}{m}} \frac{\partial}{\partial \theta}.$$

Consider the 2-dimensional surfaces in phase space defined by $P =$ constant and $\rho =$ constant. Verify that H is constant on each of these surfaces. Regarding any one of these surfaces as a manifold in its own right (a ‘submanifold’ of the whole phase space), show that $\boldsymbol{V}_H$ and $\boldsymbol{V}_P$ define independent vector fields on each surface. Convince yourself that any integral curve of $\boldsymbol{V}_H$ or $\boldsymbol{V}_P$ lies entirely within one of these surfaces.

Chapter 4

General Relativity and Gravitation

We now have at our disposal all the mathematical tools that are needed to understand the general theory of relativity and the account it offers of gravitational phenomena. Chapter 2 ended with the question ‘what is the structure of our spacetime?’ *A priori*, the possibilities are limitless: for a start, there are infinitely many dimensionalities to choose from. However, because special relativity accounts extremely well for a great many phenomena, it is clear that our spacetime must be quite similar to Minkowski spacetime. Our first task in this chapter will be to use this observation to restrict the range of possibilities that need to be considered in practice, which is more or less equivalent to adopting the principle of equivalence mentioned in chapter 2. The next step will be to find out how a given geometrical structure affects the behaviour of material objects, and this will show us how deviations of this structure from that of Minkowski spacetime can be interpreted in terms of gravitational forces. Finally, we shall investigate how the geometrical structure is determined—or at any rate influenced—by the distribution of gravitating matter and take a look at some of the phenomena that are predicted by our new theory.

4.1 The Principle of Equivalence

As we stated it in chapter 2, the principle of equivalence asserts that all gravitational effects can be eliminated within a sufficiently small region of space by adopting freely falling inertial frames of reference. Near the surface of the Earth, for example, this frame of reference is obviously accelerating relative to one fixed in the Earth and the ‘equivalence’ is between, on the one hand, the acceleration of the inertial frame relative to an earthbound observer and, on the other, the gravitational forces that appear to this observer to act on falling bodies. Let us now see what this principle asserts in terms of spacetime geometry. We shall assume that the metric tensor field $g_{\mu\nu}(x)$ with its associated metric connection (2.50) is the only geometrical structure possessed by the spacetime manifold. The square matrix formed by its components is symmetric and I shall call it $g(x)$. On transforming to a new coordinate system, the new matrix is

$$g' = \Lambda^{\mathrm{T}} g \Lambda \tag{4.1}$$

where Λ is the transformation matrix whose components were defined in (2.14) as $\Lambda^{\mu}{}_{\mu'} = \partial x^{\mu}/\partial x^{\mu'}$, and Λ^{T} is its transpose. Any symmetric matrix can be diagonalized by a transformation of this kind. Let us therefore consider a definite point P and a coordinate system in which g is diagonal at P. Assuming that none of the eigenvalues of g is zero (if one of them does vanish, then P is some sort of singular point at which odd things may happen), it will clearly be possible to adjust the scales of the coordinates so that each eigenvalue is either $+1$ or -1. If the equivalence principle is to hold in the neighbourhood of P, then the resulting $g(P)$ must be a 4×4 matrix with one eigenvalue equal to $+1$ and the other three equal to -1. Then, after renumbering the coordinates if necessary, it has the desired Minkowski-spacetime form (2.8): $g_{\mu\nu}(P) = \eta_{\mu\nu}$.

Although P can be any point, it will not in general be possible to find a coordinate system in which $g_{\mu\nu} = \eta_{\mu\nu}$ at every point. If such a coordinate system does exist then the spacetime is Minkowskian. However, it is always possible to find a coordinate system in which both $g_{\mu\nu}(P) = \eta_{\mu\nu}$ and all the first derivatives $\partial_{\sigma} g_{\mu\nu}$ vanish at P (see exercise 4.1). (Readers may like to consider in detail why there is enough freedom in coordinate transformations to achieve this, but not to diagonalize the metric at every point simultaneously.) A coordinate system of this kind may be called a *locally inertial system* at P. An observer at P who is at rest in such a system will experience the coordinate direction with the positive eigenvalue as time and the other three as spatial. According to the principle of equivalence, if the laws of physics are expressed in terms of locally inertial coordinates, they will reduce at P to the form they take in Minkowski spacetime in terms of Cartesian coordinates, and they will contain no reference to gravitational forces. This, as we are about to discover, is because gravitational forces are given by the connection coefficients (2.50), which vanish at P when expressed in locally inertial coordinates.

4.2 Gravitational Forces

Suppose for now that the metric tensor field is fixed and that it does not reduce to that of Minkowski spacetime in any system of coordinates (except locally, as discussed above). Normally, this means that spacetime is curved, and we wish to know what effect the curvature has on the laws of motion of particles. From the point of view of chapter 3, this involves finding an action appropriate to the curved spacetime. The two guiding principles here are the principle of equivalence, which we have just been discussing, and the *principle of general covariance*. In Minkowski spacetime, we concluded that the equations of motion should be covariant under Poincaré transformations because these left the metric unchanged. In curved spacetime, there are in general no coordinate transformations that leave the metric unchanged. On the other hand, any coordinate system is merely a theoretical device that enables us to label points of spacetime. The only reason for preferring a particular coordinate system would be if it permitted a specific metric tensor field to be described in an especially simple way, as is the case with Cartesian coordinates in Minkowski spacetime. If we do not commit ourselves in advance to a specific metric, then any coordinate system should be as good as any other and, in particular, equations of motion should preserve their form under any coordinate transformation. This is the meaning of general covariance.

Clearly, equations of motion will be generally covariant if they are derived from an action that is *invariant* under all transformations, namely a scalar. Scalars can be formed by contracting all the indices of any tensor with the same covariant and contravariant rank. If we allow any number of derivatives of the metric tensor field to appear in the Lagrangian, then a great many functions would be possible—for example, any function of the Ricci scalar R. In order to satisfy the principle of equivalence, however, we would like the Lagrangian to reduce to its Minkowskian form in a locally inertial frame, and our previous discussion shows that we must work only with $g_{\mu\nu}$ and its first derivatives. But to form tensors and ultimately scalars, we must use covariant derivatives rather than partial ones, and the first covariant derivative of the metric tensor field is, by definition, equal to zero (equation (2.48)). Thus, for a single particle, the Lagrangian must be a scalar formed from the vector $\dot{x}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ and the metric tensor field itself. Because of (2.44), contracting the indices of two gs gives a trivial result, and we see that the Lagrangian can only be a function of the scalar quantity $X = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$. As in Minkowski spacetime, we find that the detailed form of this function is immaterial, and we need only replace $\eta_{\mu\nu}$ in (3.32) by $g_{\mu\nu}$:

$$S = -\frac{1}{2}m\int \mathrm{d}\tau\, g_{\mu\nu}(x(\tau))\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}. \tag{4.2}$$

The equation of motion for a free particle moving in the curved spacetime is the Euler–Lagrange equation obtained by varying (4.2) with respect to the path $x^\mu(\tau)$, namely

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(g_{\mu\nu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\right) - \frac{1}{2}g_{\sigma\nu,\mu}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} = 0. \tag{4.3}$$

As in chapter 2, the comma before the index μ is a shorthand for $\partial/\partial x^\mu$. After carrying out the differentiation and raising the non-contracted index, this may be written as

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau} = 0 \tag{4.4}$$

which is the equation of a geodesic curve, introduced in chapter 2 as the curved-space analogue of a straight line. The affine connection coefficients are those given by (2.50).

If our qualitative discussions of the relativistic theory of gravity are to stand up, it must now be possible to find a set of circumstances under which (4.4) can be reinterpreted as the equation of a particle moving through Minkowski or Galilean spacetime under the influence of a gravitational field. I shall now show what these circumstances are. An obvious requirement is that the metric should be only slightly different from the $\eta_{\mu\nu}$ of Minkowski spacetime, so let us write it as

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{4.5}$$

where $h_{\mu\nu}$ is a small correction. If we keep only terms of first order in $h_{\mu\nu}$, then the connection coefficients are

$$\Gamma^{\mu}_{\ \nu\sigma} = \tfrac{1}{2}\eta^{\mu\lambda}(h_{\lambda\nu,\sigma} + h_{\lambda\sigma,\nu} - h_{\nu\sigma,\lambda}) + \mathrm{O}(h^2). \tag{4.6}$$

The second requirement is that the particle should be moving, relative to our chosen coordinate system, very slowly compared with the speed of light. This is normally true in those practical situations that appear to support the Newtonian account of gravity; for example, the orbital speed of the Earth around the sun is about $10^{-4}c$. The element of proper time along the particle's path is given, according to (4.5), by $c^2\mathrm{d}\tau^2 = (\eta_{\mu\nu}+h_{\mu\nu})\mathrm{d}x^\mu\mathrm{d}x^\nu$ and since, for a slowly moving particle, $\mathrm{d}\boldsymbol{x}/\mathrm{d}\tau$ is negligible compared with $\mathrm{d}t/\mathrm{d}\tau$, we have approximately

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} \simeq (1+h_{00})^{-1/2} \simeq 1 - \tfrac{1}{2}h_{00}. \tag{4.7}$$

By the same token, the spatial components of (4.4) can be written (using the convention that Latin indices $i, j, k, \ldots$ denote spatial directions and recalling that $x^0 = ct$) as

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2} + \Gamma^{i}_{\ 00}c^2\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 \simeq 0. \tag{4.8}$$

The final requirement is that the variation with time of the metric tensor field and hence, as we shall see immediately, of the gravitational field is negligible. This has two consequences. First, $\mathrm{d}t/\mathrm{d}\tau$ in (4.7) is approximately a constant, so in (4.8) we can set $\mathrm{d}^2x^i/\mathrm{d}\tau^2 \simeq (\mathrm{d}t/\mathrm{d}\tau)^2\mathrm{d}^2x^i/\mathrm{d}t^2$ and $(\mathrm{d}t/\mathrm{d}\tau)^2$ cancels out. Second, terms in the connection coefficients which involve time derivatives can be neglected. In particular, the coefficient that appears in (4.8) is just

$$\Gamma^{i}_{\ 00} \simeq \tfrac{1}{2}h_{00,i}. \tag{4.9}$$

So, on multiplying (4.8) by the mass of the particle, we get

$$m\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2} \simeq -m\frac{\partial}{\partial x^i}V \tag{4.10}$$

where V is the gravitational potential of the Newtonian theory, now to be identified as

$$V = \tfrac{1}{2}c^2h_{00}. \tag{4.11}$$

At this point, then, our mathematical account of spacetime geometry begins to make contact with actual observations. If the above requirements are met, we say that the *Newtonian limit* applies. In this limit, we can pretend that Minkowskian or Galilean geometry is correct. The small error that we incur by doing this is detectable by virtue of the gravitational force on the right-hand side of (4.10), which is related to the true metric through (4.11). Of course, we are not really entitled yet to identify the V in these equations as a *gravitational* potential, rather than a potential of some other kind. We have, certainly, obtained one of the hallmarks of gravity, namely that the force in (4.10) is proportional to the inertial mass of the test particle. The other half of the story is that V should be of the correct form. For example, in the neighbourhood of the Earth, V should be approximately equal to $-GM/r$, where G is Newton's constant, M the mass of the Earth and r the distance from its centre. In the next section, we shall see how this comes about.

4.3 The Field Equations of General Relativity

We have come some way towards answering the question 'what is the structure of our spacetime?'. On empirical grounds, we have seen that it cannot be too far removed from that of Minkowski spacetime. Moreover, we have seen how small deviations from the Minkowski metric can be interpreted in terms of a force field that we would like to identify with gravity. Our basic assumption will now be that the metric tensor field is a physical object whose behaviour is governed, like that of other physical objects, by an action principle. Although gravity is properly viewed as an 'apparent' force, which disappears when we adopt a truly inertial frame of reference, it is helpful to some extent to think of gravity by analogy with electromagnetism. Thus, the action (4.2), with the metric tensor field decomposed as in (4.5), may be thought of as analogous to the first two terms of (3.55). These lead to the equation of motion (3.56) or (3.58) (analogous to (4.4) or (4.10)) of a charged particle in the presence of given electric and magnetic fields. To find out what electric and magnetic fields are actually present, we have to solve Maxwell's equations (3.52), which relate derivatives of the fields on the left-hand side to the charge density and currents on the right-hand side. To derive Maxwell's equations, we require the final term in (3.55), which depends on the electromagnetic fields alone.

To find out what the metric tensor field is, for a given region of space containing a given distribution of matter, we must solve the gravitational analogues of Maxwell's equations. These are *Einstein's field equations*. The currents on the right-hand side will turn out to be the stress tensor given in (3.42). The left-hand side, analogous to $\partial_\mu F^{\mu\nu}$ in (3.52), is the *Einstein curvature tensor*, which is constructed from the metric tensor field in a manner we have yet to discover. To do this, we must evidently add to the action a term analogous to the last term of (3.55). It must be a scalar quantity, containing just the metric tensor field and its derivatives.

There is one mathematical detail to be sorted out first. Namely, we need to know how to integrate over spacetime in a covariant manner. Suppose, to take the simplest case, that we have a coordinate system in which the metric tensor field at the point x is diagonal with elements g_{00}, g_{11}, g_{22} and g_{33}. An infinitesimal time interval is $\mathrm{d}t = c^{-1}(g_{00})^{1/2}\mathrm{d}x^0$ and infinitesimal distances are $\mathrm{d}x = (-g_{11})^{1/2}\mathrm{d}x^1$, etc. Therefore, the infinitesimal spacetime volume element is

$$\mathrm{d}(\text{spacetime volume}) = c^{-1}\mathrm{d}^4x\,(-g(x))^{1/2} \tag{4.12}$$

where $g(x)$ denotes the determinant of the metric tensor field. On transforming to a new coordinate system, d^4x is multiplied by a Jacobian factor, which is the determinant of the transformation matrix (2.13). Readers should have no difficulty in verifying that this is exactly cancelled by the determinant of the inverse matrix that transforms $g(x)$ according to (4.1). Thus, the volume element (4.12) is a scalar, retaining the same form in all coordinate systems. Correspondingly, we may define a scalar δ function

$$(-g(x))^{-1/2}\delta^4(x-y) \tag{4.13}$$

which has the desired properties when used in conjunction with the scalar volume element (4.12).

Beyond the requirement that the geometrical contribution to the action should be a scalar, there seems to be no *a priori* way of knowing what form it should take. Arguably, the form that has been found to work is the simplest possible one, but simplicity is a somewhat subjective and ill-defined criterion. It also has the feature that the resulting equation of motion for $g_{\mu\nu}$, like those for other physical quantities, contains only first and second derivatives of $g_{\mu\nu}$, but it is not altogether clear that this need be insisted on. At any rate, the standard version of general relativity is obtained by taking the total action to be

$$S = \int \mathrm{d}^4x \left[\mathcal{L}_{\text{matter}}(x) + \mathcal{L}_{\text{grav}}(x)\right] \tag{4.14}$$

where the Lagrangian densities for matter and for gravitational fields are

$$\mathcal{L}_{\text{matter}}(x) = -\tfrac{1}{2}\sum_n m_n \int \mathrm{d}\tau_n \delta^4\Big(x - x_n(\tau_n)\Big)\, g_{\mu\nu}(x)\dot{x}_n^{\mu}(\tau_n)\dot{x}_n^{\nu}(\tau_n) \tag{4.15}$$

$$\mathcal{L}_{\text{grav}}(x) = -\frac{1}{c\kappa}(-g(x))^{1/2}\left[\Lambda + \tfrac{1}{2}R(x)\right]. \tag{4.16}$$

By integrating $\mathcal{L}_{\text{matter}}(x)$ over all spacetime, we get a term of the form (4.2) for each particle of matter. Notice that the factors of $(-g(x))^{1/2}$ have cancelled between the spacetime volume element and the δ function. In $\mathcal{L}_{\text{grav}}(x)$, $R(x)$ is the Ricci curvature scalar (2.51) and Λ is a constant, called the *cosmological* *constant*. The overall constant κ determines the strength of the coupling between geometry and matter, and consequently the strength of gravitational forces. It must obviously be related to Newton's constant, and we shall shortly derive the exact relationship.

By requiring the action (4.14) to be stationary against variations in each of the particles trajectories, we obtain an equation of motion of the form (4.4) for each particle. The field equations are obtained by requiring it to be stationary against variations in $g_{\mu\nu}(x)$. In principle, this is no more difficult than obtaining Maxwell's equations from (3.54), but the algebra is considerably more involved. Exercise 4.2 offers guidelines for carrying the calculation through, but here I shall just quote the result: *Einstein's field equations* are

$$R^{\mu\nu} - \left(\tfrac{1}{2}R + \Lambda\right) g^{\mu\nu} = \kappa T^{\mu\nu}. \tag{4.17}$$

The two terms $G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu}$ (in which $R^{\mu\nu}$ is the Ricci tensor (2.36) with its indices raised) constitute what is sometimes called the *Einstein curvature tensor*. The cosmological constant Λ is, according to the best astronomical evidence, very close to zero in our universe and may generally be omitted. At the time of writing, there is no understanding of why Λ should be close or equal to zero (though many people have attempted speculative explanations), and indeed this question is widely regarded as one of the most important mysteries remaining in modern cosmology. The stress tensor on the right-hand side is

$$T^{\mu\nu}(x) = \frac{c}{(-g(x))^{1/2}} \sum_n \int \mathrm{d}\tau_n\, m_n \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau_n}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau_n}\, \delta^4(x - x_n(\tau_n)). \tag{4.18}$$

It differs from the Minkowski-spacetime tensor (3.42) only insofar as the invariant δ function (4.13) has been used.

If the relativistic theory of gravity is to work, it must now be possible to show that the potential $V(x)$ defined in (4.11) reduces to the Newtonian potential in the appropriate limit. The Newtonian potential of a point mass M at a distance r from it is $V(r) = -GM/r$. Equivalently (as is shown in any textbook on electricity for the analogous Coulomb potential), for a static mass distribution of density $\rho(\boldsymbol{x})$, the potential satisfies Poisson's equation

$$\nabla^2 V = 4\pi G\rho. \tag{4.19}$$

I shall show that this equation follows, in the Newtonian limit, from the $(0, 0)$ component of the field equations (4.17). To this end, it is convenient to rewrite these equations in the following way. First, define the scalar quantity T by $T = g_{\mu\nu}T^{\mu\nu}$. By contracting (4.17) with $g_{\mu\nu}$, we find that $R = -4\Lambda - \kappa T$ and on substituting this back into (4.17) we get the alternative version

$$R^{\mu\nu} = \kappa(T^{\mu\nu} - \tfrac{1}{2}Tg^{\mu\nu}) - \Lambda g^{\mu\nu}. \tag{4.20}$$

Now assume that a coordinate system can be found in which the matter giving rise to the gravitational potential is at rest and in which the metric tensor field is close to that of Minkowski spacetime, as in (4.5). To the order of accuracy we require, the right-hand side of (4.20) can be evaluated with $h_{\mu\nu} = 0$. For particles at rest, we have $\mathrm{d}x^\mu/\mathrm{d}\tau = (c, 0, 0, 0)$, and this can be used in (3.42) to find the stress tensor. The density is expressed by the $\mu = 0$ component of (3.41) when A is taken to be the mass of a particle, and we find that all components of the stress tensor are zero except for $T^{00} = \rho c^2$, so that $T = \rho c^2$ also. (This also agrees with (3.43), bearing in mind that the symbol ρ in that equation is the energy density, whereas here I am using it to stand for the mass density.) In the Newtonian limit discussed in the last section, the (0, 0) component of the Ricci tensor field is given approximately by

$$R^{00} \simeq \tfrac{1}{2}\sum_{i=1}^{3} \partial_i \partial_i h_{00}. \tag{4.21}$$

With h_{00} identified as in (4.11), the (0, 0) component of (4.20) now reads

$$\nabla^2 V = \left(\tfrac{1}{2}\kappa\rho c^2 - \Lambda\right) c^2. \tag{4.22}$$

This is identical with Poisson's equation (4.19) provided that the cosmological constant is negligibly small and that we identify the constant κ as

$$\kappa = 8\pi G/c^4. \tag{4.23}$$

Equations (4.4) and (4.17) constitute the general-relativistic theory of gravity. So long as we have values for the two constants κ and Λ, these equations may in principle be applied to any specific physical situation, their solutions yielding predictions that can be tested against actual observations. The value of κ is determined experimentally by (4.23), but the cosmological constant is, as mentioned above, rather more puzzling. In Einstein's original formulation of the theory, it was zero—which is to say that it did not appear at all. For most purposes, it is assumed to be zero, and this leads to a number of well-verified predictions, some of which are discussed in the following section and in chapter 14.

The extent of our knowledge of the actual value of Λ is that it cannot be large enough to invalidate these predictions. (At the time of writing, there is some evidence for an acceleration of the expansion of the universe that might be explained by a small, nonzero value of Λ, but this cannot yet be taken as reliable.) In (4.22) the quantity $\Lambda/\kappa c^2 = \Lambda c^2/8\pi G$ appears as a negative 'mass density of the vacuum', to be considered along with the density of real matter. This is a somewhat dangerous observation, because Λ appears in other places as well. (For example, in (14.18) its net effect is equivalent to a *positive* mass density.) Nevertheless, a rough and ready method of placing upper bounds on the value of $\Lambda/\kappa c^2$ is to argue that it must be significantly smaller than the average density of a system that is well described by the theory with $\Lambda = 0$. For example, the solar system is described by this theory to within the accuracy of observations and of the approximations needed to obtain numerical theoretical predictions. A suitable 'density' might be the mass of the Sun divided by the volume of a sphere that just encloses the orbit of Pluto, which gives about $3 \times 10^{-12}\,\mathrm{g\,cm^{-3}}$, and the agreement of theory with experiment would be upset if the vacuum density were comparable with this. Applying the same argument to much larger systems such as clusters of galaxies (which are much less precisely understood than the solar system), we obtain a limit of the kind

$$\Lambda c^2/8\pi G \lesssim 10^{-29}\,\mathrm{g\,cm^{-3}}. \tag{4.24}$$

This is roughly the average density of observable matter in the universe and is, of course, vastly smaller than the densities of familiar materials. Whether it is small in an absolute sense depends on our finding some fundamental quantity with the dimensions of a density with which to compare it. We shall see later that such a comparison can be made, which suggests that the smallness of Λ is even more striking than the number quoted in (4.24).

4.4 The Gravitational Field of a Spherical Body

To find out how the general-relativistic theory of gravity differs from the Newtonian one, we must, of course, find exact solutions to (4.4) and (4.17), or at least approximate solutions that go beyond the Newtonian approximation. I shall illustrate the nature of general-relativistic effects by considering Schwarzschild's solution of the field equations for the metric tensor field associated with a massive spherical body and some of its elementary consequences.

4.4.1 The Schwarzschild solution

The task of finding a general solution to the field equations is too difficult to contemplate, and it is usually possible to find particular solutions only when symmetry or other requirements can be used to reduce the 10 independent components of the metric tensor field to a more manageable number. The solution found by Schwarzschild (1916), although it is an exact solution, rests on several simplifying assumptions. First, we ask for the gravitational field of a spherically symmetric body and assume that the metric will also be spherically symmetric. Second, since we anticipate that gravitational effects will be extremely weak at large distances from the body, the metric should approach that of Minkowski spacetime at large distances. We therefore use polar coordinates (t, r, θ, ϕ) and expect that for large r the line element will be approximately

$$c^2\mathrm{d}\tau^2 \simeq c^2\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right). \tag{4.25}$$

It must be borne in mind that these coordinates cannot necessarily be interpreted as time, radial distance and angles in the elementary sense, although these interpretations should become valid in the large r region where (4.25) is valid.

The final assumption is that, in these coordinates, the components of the metric tensor field are independent of the coordinate t. This implies, in particular, that an observer in the large r region will see a static gravitational field. As a matter of fact, the only assumption which is really needed is that of spherical symmetry. There is a theorem due to G D Birkhoff (explained, for example, by Weinberg (1972)), which shows that the only spherically-symmetric solution for the metric of a spacetime that is empty apart from a central spherical body is the time-independent Schwarzschild solution. Here, to make matters simpler, I shall take it as an extra assumption that the metric is static. With these assumptions, the line element can be written as

$$c^2\mathrm{d}\tau^2 = A(r)c^2\mathrm{d}t^2 - B(r)\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right). \tag{4.26}$$

The two functions $A(r)$ and $B(r)$, which should approach the value 1 for large r, remain to be determined. A third unknown function $C(r)$ could have been included in the coefficient of the angular term. However, we could then define a new radial coordinate by $r'^2 = C(r)r^2$, and so recover the form (4.26) with A and B appropriately redefined.

We shall consider only the *exterior* solution, namely the metric as it exists outside the central body. In this region, there is no matter, so, taking the cosmological constant to be zero, we have to solve (4.17) in the special case that $\Lambda = T^{\mu\nu} = 0$. This is actually a set of ten equations for the ten independent components of the metric tensor field. Provided, as is in fact the case, that our assumptions are consistent with the structure of the field equations, it will be possible to find functions $A(r)$ and $B(r)$ such that all ten equations are satisfied. The task of finding these functions and verifying that all the field equations are satisfied is straightforward, but quite lengthy, although the result is a simple one. I shall outline the steps and leave it to sufficiently energetic readers to fill in the details. The components $g_{\mu\nu}$ can be read off from (4.26). We must use them to calculate the connection coefficients (2.50) and thence the Ricci tensor (2.36) and the scalar curvature (2.51). A useful short cut to finding the connection coefficients is to write out the action (4.2) explicitly:

$$S = -\tfrac{1}{2}m\int \mathrm{d}\tau\left[c^2A(r)\dot{t}^2 - B(r)\dot{r}^2 - r^2\left(\dot{\theta}^2 + \sin^2\theta\,\dot{\phi}^2\right)\right]. \tag{4.27}$$

By varying each of the coordinates, it is easy to find the Euler–Lagrange equations, from which the $\Gamma^{\mu}_{\nu\sigma}$ can be picked out by comparison with (4.4).

There is now nothing for it but to work out the components of $R^{\mu\nu}$ and equate them to zero. (By contracting $R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} = 0$ with $g_{\mu\nu}$, we find that both $R^{\mu\nu}$ and R must vanish.) As it turns out, all the off-diagonal elements vanish identically. The remaining four equations are differential equations for $A(r)$ and $B(r)$, which have the solution $A(r) = 1/B(r) = 1 + \alpha/r$, where α is a constant of integration. To identify the constant, we note that h_{00} in (4.11) is just α/r. For large r, this is indeed small and must equal $2/c^2$ times the Newtonian potential $-GM/r$, where M is the mass of the central body. The Schwarzschild line element is therefore

$$c^2 d\tau^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2 r}\right)^{-1} dr^2 - r^2 \left(d\theta^2 + \sin^2\theta \, d\phi^2\right). \tag{4.28}$$

It has an obvious peculiarity at the *Schwarzschild radius*

$$r_S = 2GM/c^2 \tag{4.29}$$

which has, for example, values of 0.886 cm for the Earth, 2.95 km for the Sun and 2.48×10^{-52} cm for a proton. As we shall see, this singularity is associated with the possibility of 'black holes'. Remember, however, that (4.28) is the exterior solution for the metric, valid outside the massive body. It does not follow that there is a black hole of radius 0.886 cm lurking at the centre of the Earth! Before discussing this in more detail, we shall take a look at some more prosaic features of the Schwarzschild solution.

4.4.2 Time near a massive body

A normal body, such as the Earth or the Sun, is larger than the Schwarzschild radius calculated from its mass. Let us consider a stationary observer near such a body to be one whose (r, θ, ϕ) coordinates are fixed. For such an observer, the flow of proper time is measured by

$$d\tau = \left(1 - \frac{r_S}{r}\right)^{1/2} dt \tag{4.30}$$

as we discover by setting $dr = d\theta = d\phi = 0$ in (4.28). The time experienced by a stationary observer is thus proportional to the coordinate t, but with a factor that changes with r. Two events occurring at the same value of t will appear simultaneous to any stationary observer, and therefore the spacetime can be separated a meaningful way into three-dimensional spatial slices, each labelled by its own value of t. All stationary observers agree on this splitting, but the time that elapses between two given values of t is different for observers at different radial positions.

The variation of time intervals with radial position can be investigated by the shift it causes in atomic spectral lines. Consider a radiating atom located at $r_{\rm at}$ and an observer at $r_{\rm obs}$. Suppose a pulse of light is emitted at coordinate time $t_{\rm e}$ and received at $t_{\rm r}$, and a second pulse is emitted at $t_{\rm e} + \Delta t_{\rm e}$, being received at $t_{\rm r} + \Delta t_{\rm r}$ (see figure 4.1). Since the metric is independent of t, the paths of the two pulses through spacetime are exactly similar, and therefore the coordinate time interval $t_{\rm r} - t_{\rm e}$ between emission and reception of the first pulse is equal to the corresponding interval $(t_{\rm r} + \Delta t_{\rm r}) - (t_{\rm e} + \Delta t_{\rm e})$ for the second. It follows that the coordinate time interval $\Delta t_{\rm e}$ between the moments when the two pulses are emitted is equal to the interval $\Delta t_{\rm r}$ between the moments at which they are

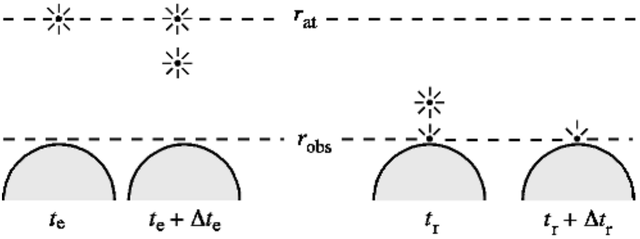

$r_{\rm at}$
$r_{\rm obs}$
$t_{\rm e}$
$t_{\rm e} + \Delta t_{\rm e}$
$t_{\rm r}$
$t_{\rm r} + \Delta t_{\rm r}$

Figure 4.1. Passage of two pulses of light from a radiating atom to an observer in the gravitational field of a spherical body.

received: $\Delta t_{\rm e} = \Delta t_{\rm r}$. The corresponding proper time intervals are therefore different, and the ratio of the observed frequency of the received wave to the frequency of the wave as emitted by the atom follows trivially from (4.30):

$$\frac{\text{observed frequency}}{\text{frequency at emission}} = \frac{(\Delta\tau_{\rm obs})^{-1}}{(\Delta\tau_{\rm at})^{-1}} = \left(\frac{1 - r_{\rm S}/r_{\rm at}}{1 - r_{\rm S}/r_{\rm obs}}\right)^{1/2}. \tag{4.31}$$

This ratio involves only the $(0, 0)$ component of the metric tensor, which we have identified in terms of the gravitational potential. In general, for a static spacetime (that is, for one that can be divided into identical spatial slices), we have

$$\frac{\text{observed frequency}}{\text{frequency at emission}} = \left(\frac{1 + 2V_{\rm at}/c^2}{1 + 2V_{\rm obs}/c^2}\right)^{1/2}. \tag{4.32}$$

In a weak gravitational field, the frequency shift $\Delta\nu = \nu_{\rm obs} - \nu_{\rm at}$ is given approximately by

$$\frac{\Delta\nu}{\nu} = \frac{V_{\rm at} - V_{\rm obs}}{c^2}. \tag{4.33}$$

Although this shift can have either sign, what can normally be observed in practice is light from the atmospheres of stars. The radiating atom in this case is at a lower gravitational potential than an earthbound telescope, so a *gravitational redshift* is observed. Such observations confirm the prediction (4.32) to precisions of a few percent. A method of measuring frequency shifts in the Earth's gravitational field was devised by Pound and Rebka (1960), who used the Mössbauer effect to determine the change in frequency of γ rays from ^{57}Fe nuclei on travelling a vertical distance of some 22 m. In this case, the frequency shift can be deduced from a simple application of the equivalence principle, without the full machinery of general relativity (see exercise 4.3).

4.4.3 Distances near a massive body

Within an equal-time slice of the Schwarzschild spacetime, distances are measured by the spatial part of the line element

$$\mathrm{d}l^2 = \left(1 - \frac{r_\mathrm{S}}{r}\right)^{-1} \mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right). \tag{4.34}$$

This is a non-Euclidean space, and the departure from Euclidean geometry may be illustrated by the fact that the circumference of a circle is not equal to 2π times its radius. Consider a circle concentric with the central body in the equatorial plane $\theta = \pi/2$ at a fixed radial coordinate r. Its circumference is

$$\text{circumference} = \int_0^{2\pi} \frac{\mathrm{d}l}{\mathrm{d}\phi}\,\mathrm{d}\phi = 2\pi r. \tag{4.35}$$

Its radius cannot be determined exactly, because (4.34) is valid only outside the central body. We can, however, compare two circles of coordinate radii r_1 and r_2. In Euclidean geometry, the difference between their circumferences is 2π times the difference between their radii. In the Schwarzschild space, the difference in circumference is $2\pi(r_2 - r_1)$, but the radial distance between them is

$$\text{radial distance} = \int_{r_1}^{r_2} \frac{\mathrm{d}l}{\mathrm{d}r}\mathrm{d}r = \int_{r_1}^{r_2} \frac{\mathrm{d}r}{(1 - r_\mathrm{S}/r)^{1/2}} = r_2 f(r_2) - r_1 f(r_1) \tag{4.36}$$

where the function $f(r)$ is

$$f(r) = \left(1 - \frac{r_\mathrm{S}}{r}\right)^{1/2} + \left(\frac{r_\mathrm{S}}{r}\right)\ln\left\{\left(\frac{r}{r_\mathrm{S}}\right)^{1/2}\left[1 + \left(1 - \frac{r_\mathrm{S}}{r}\right)^{1/2}\right]\right\}. \tag{4.37}$$

When r is much greater than r_S, this may be approximated as

$$f(r) \simeq 1 + \left(\frac{r_\mathrm{S}}{r}\right)\ln\left[2\left(\frac{r}{r_\mathrm{S}}\right)^{1/2}\right] \tag{4.38}$$

and for two circles satisfying this condition, we find

$$\frac{\text{difference in circumference}}{\text{radial distance}} \simeq 2\pi\left[1 - \frac{1}{2}\left(\frac{r_\mathrm{S}}{r_2 - r_1}\right)\ln\left(\frac{r_2}{r_1}\right)\right] \tag{4.39}$$

provided that $r_2 - r_1$ is also larger than r_S. As an example, if r_S is the Schwarzschild radius of the Sun, r_1 is the radius of the Sun (6.96×10^8 m) and r_2 is the semi-latus rectum of the orbit of Mercury (5.5×10^{10} m), then the correction term is about 10^{-7}. For many purposes, therefore, the solar system can adequately be described in terms of Euclidean geometry.

4.4.4 Particle trajectories near a massive body

The analogy drawn above between the field equations and Maxwell's equations may be misleading in one important respect: the field strength tensor (3.51) is linear in the electromagnetic fields, while the curvature tensors are nonlinear in the metric tensor field. Suppose, for example, that we wish to calculate, according to classical mechanics, the orbit of an electron near a positive nucleus, which we take to remain stationary. The linearity of the field strength tensor allows us to express the total electric field as the sum of fields due to the nucleus and the electron. The field due to the electron exerts no force on the electron itself. It can be subtracted from the total field, and we simply regard the electron as moving in the field of the nucleus. In general, this cannot be done with gravity. Given, say, a star and a single planet, the true metric cannot be expressed as the sum of two Schwarzschild metrics. If we wish to find the metric and the relative motion of the two bodies, it is necessary to solve the whole problem in one go: since we do not know the metric, we cannot immediately find the orbits and, not knowing these, we cannot write down any explicit form for the stress tensor that appears in the field equations we must solve for the metric. In fact, the exact solution of this two-body problem is not known.

What we can do without too much trouble is to work out the trajectories of 'test particles' in the Schwarzschild spacetime—or at least we can write down their equations of motion and solve these by some approximate means. A test particle is one whose effect on the metric is negligible, and its equations of motion are the geodesic equations (4.4) with the connection coefficients calculated in this case from the Schwarzschild metric. I shall write out explicitly only the form of these equations that applies to motion in the equatorial plane: this can, of course, be any plane passing through the centre of the massive body if we choose our coordinates appropriately. With θ fixed at $\pi/2$, the equations are

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\left(1-\frac{r_\mathrm{S}}{r}\right)\dot{t}\right]=0 \tag{4.40}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(r^2\dot{\phi}\right)=0 \tag{4.41}$$

$$\left(1-\frac{r_\mathrm{S}}{r}\right)^{-1}\ddot{r}+\frac{1}{2}c^2\left(\frac{r_\mathrm{S}}{r^2}\right)\dot{t}^2-\frac{1}{2}\left(1-\frac{r_\mathrm{S}}{r}\right)^{-2}\left(\frac{r_\mathrm{S}}{r^2}\right)\dot{r}^2-r\dot{\phi}^2=0. \tag{4.42}$$

As in previous equations, the overdot denotes $\mathrm{d}/\mathrm{d}\tau$.

The derivation of the equations of motion (4.4) was valid for massive particles. For photons, or other massless particles, the action (4.2) vanishes. To deal with this case, we simply define a new parameter λ such that $\mathrm{d}\tau = m\mathrm{d}\lambda$. The mass then disappears from the action and can be set to zero. The equations of motion (4.4) then follow as before, but with τ replaced by λ. The trajectories for massless particles are still geodesics, but are not parametrized by proper time. Clearly, indeed, they are *null geodesics*, along which $\mathrm{d}\tau = 0$.

These equations lead to a number of interesting predictions when applied to the solar system. Light passing close to the Sun is predicted to be deflected by 1.75 seconds of arc, and the expeditions of Dyson, Eddington and Davidson to observe this effect during a total eclipse in 1919 resulted in one of the earliest confirmations of Einstein's theory. (Their measurements were actually not precise enough to justify the confirmation that was claimed at the time, but later, more accurate observations do confirm the theoretical result.) When the planets are treated as test particles, it is found that their orbits are not elliptical as in the simple Newtonian theory, but can be described as ellipses whose perihelia (points of closest approach to the Sun) precess slowly. The largest precession rate, that for Mercury, is predicted to be some 43 seconds of arc per century. This is also in agreement with observations, but only when the perturbing effect of other planets is taken into account. Planetary orbits have, of course, been studied for centuries and are known with great precision. Even within Newtonian theory, the approximation of treating the planets as test particles is far too crude, and their perturbing influence on each other must be taken into account. These perturbations themselves cause precessions, to which the general-relativistic effect is a small correction. In order to apply general relativity to the solar system in a meaningful way, systematic methods of obtaining corrections to the detailed Newtonian theory must be devised. These techniques, known as *post-Newtonian approximations* are discussed in specialized textbooks, but are well beyond the scope of this one. Finally, as first worked out by Shapiro (1964), radar signals reflected from a neighbouring planet are slightly delayed by comparison with their round-trip time according to the Newtonian theory. The simpler aspects of these phenomena are explored in the exercises.

4.5 Black and White Holes

So far, we have considered the spacetime near a massive body whose radius is larger than its Schwarzschild radius r_{S}. In this section, we shall consider the case of an object that is smaller than its Schwarzschild radius. First, let us see whether it is possible to make sense of the metric (4.28) all the way down to $r = 0$. This metric is valid only outside the central body, so physically we will want to know what has happened to the said body. This question will be addressed in due course: for now, let us take it to be an idealized point particle, which nevertheless has a substantial mass M.

To simplify matters, I shall discuss only the paths of free particles moving in the radial direction, which are described by the two functions $r(\tau)$ and $t(\tau)$. Remember that while τ is by definition the time experienced by the particle, the coordinates r and t have no unique interpretation as distances or times. In the region where r is large, however, they are, to a good approximation, the radial distance and time as experienced by a stationary observer. The paths of radially moving particles are most easily found by using the equation

$$\dot{r}^2 = \left(1 - \frac{r_S}{r}\right)^2 c^2 \dot{t}^2 - c^2 \left(1 - \frac{r_S}{r}\right) \tag{4.43}$$

which follows from the line element (4.28) with $\mathrm{d}\theta = \mathrm{d}\phi = 0$. Eliminating $\dot{t}$ between this equation and (4.42) with $\dot{\phi} = 0$, we find the radial equation of motion

$$\ddot{r} = -\frac{c^2 r_S}{2r^2}. \tag{4.44}$$

In view of the definition (4.29) of r_S, this is precisely the equation satisfied by a particle in the Newtonian potential $V = -GM/r$. Two particular solutions are those in which the particle passes through the point r_0 at time $\tau = 0$ with the corresponding escape velocity $v_{\mathrm{esc}} = (2GM/r_0)^{1/2} = c(r_S/r_0)^{1/2}$, in either the outward or the inward direction. They are

$$r(\tau) = \left(r_0^{3/2} \pm \tfrac{3}{2} c r_S^{1/2} \tau\right)^{2/3} \tag{4.45}$$

where the positive sign corresponds to an outgoing particle and the negative sign to an ingoing one. In either case, the particle can apparently pass through the point $r = r_S$ without encountering anything unusual.

Suppose that r_0 is greater than r_S. The solution for $t(\tau)$ is most easily obtained by (i) expressing $t(\tau)$ as a function of $r(\tau)$, so that $\dot{t} = \dot{r}\,\mathrm{d}t/\mathrm{d}r$ and (ii) noting that, for the particular solution (4.45), we have $\dot{r}^2 = c^2 r_S/r$. Making these substitutions in (4.43), we find

$$c r_S^{1/2} \frac{\mathrm{d}t}{\mathrm{d}r} = \pm \frac{r^{3/2}}{r - r_S} \tag{4.46}$$

which can be integrated to give

$$ct = \pm r_S^{-1/2} \left\{ \tfrac{2}{3} r^{3/2} + 2 r_S r^{1/2} + r_S^{3/2} \ln\left[\frac{(r/r_S)^{1/2} - 1}{(r/r_S)^{1/2} + 1} \right] \right\}. \tag{4.47}$$

We could add a constant of integration to specify the time at which the particle passes through r_0, but this is of no great interest. We see that, as an ingoing particle approaches r_S, its coordinate $t(\tau)$ approaches $+\infty$, although the proper time interval that it experiences while travelling from r_0 to r_S is finite, being equal to $2(r_0^{3/2} - r_S^{3/2})/3cr_S^{1/2}$. This means that in the neighbourhood of r_S, the coordinate t is no longer useful as a measure of physical time. Correspondingly, the metric given by (4.28) does not give a useful description of the geometry near r_S, because one of its components becomes infinite.

Although we have done the calculation only for one special kind of particle trajectory, much the same thing happens for any trajectory passing through r_S. Mathematically, we have to say that the spacetime manifold on which the metric

(4.28) is valid does not include the spherical surface $r = r_S$. Strictly speaking, this metric applies to two separate spacetimes, namely the two regions $r > r_S$ and $r < r_S$. In that case, what becomes of our particle when it reaches the edge of the first region, in which it started? There are two possibilities. One is that the singularity at $r = r_S$ is a genuine singularity of the geometrical structure. If so, then the particle would have reached the end of the spacetime available to it. We would have to investigate whether it could be reflected, remain trapped on the 'edge of the universe' or simply disappear from the universe altogether. In view of the fact that its radial coordinate (4.45) passes perfectly smoothly through r_S, it seems unlikely that such measures should be necessary. The other possibility is that the singularity is merely a 'coordinate singularity'. That is to say, the particle has not reached the end of spacetime, but merely the end of that part of spacetime for which t serves as a useful coordinate. This second possibility is in fact the correct one. Nevertheless, from a mathematical point of view, we have at hand only the region $r > r_S$. We must add on to it a second region, in which $r < r_S$, which is an extension of the same geometrical structure. This will be possible if we can trade in t for a new coordinate which will describe a smooth join between the two regions. This means that when we express the line element (4.28) in terms of the new coordinate, all the components of the metric tensor field will be smooth at r_S.

Let us call the region $r > r_S$ region I. This region covers most of the universe, although it is a universe populated only by 'test particles' and therefore cannot describe the whole of our actual universe. Region I has in fact two 'edges' at $r = r_S$ and $t = +\infty$ or $t = -\infty$. At these two edges, we can join on two new regions. That which joins on at $t = +\infty$, called region II, is the one into which ingoing particles fall; that which joins on at $t = -\infty$, called region II$'$ is one from which outgoing particles can emerge. Each of these regions has the same geometrical structure as the region $r < r_S$ of the original Schwarzschild solution; the trick is to find a way of smoothly joining the various regions together. The join between regions I and II can be described in terms of the Eddington–Finkelstein coordinate v, defined by

$$v = ct + r + r_S \ln\left(\frac{r}{r_S} - 1\right). \tag{4.48}$$

If we substitute for t the expression (4.47) with the $-$ sign to represent the path of an ingoing particle, we see that v remains finite as the particle passes through r_S. Moreover, when written in terms of v, the line element becomes

$$c^2 \mathrm{d}\tau^2 = \left(1 - \frac{r_S}{r}\right) \mathrm{d}v^2 - 2\mathrm{d}v\mathrm{d}r \tag{4.49}$$

which is perfectly smooth at the boundary between regions I and II. To describe the boundary with region II$'$, we can use instead the coordinate w, defined by

$$w = ct - r - r_S \ln\left(\frac{r}{r_S} - 1\right) \tag{4.50}$$

in terms of which the line element takes the form of (4.49) with dv replaced by $-$dw.

The boundary between regions I and II can be crossed only by ingoing particles and, in fact, only by ingoing light rays also. Nothing ever crosses from region II into region I, for which reason region II is called a *black hole*. Conversely, particles and light rays may cross from region II′ into region I, but not in the opposite direction, so region II′ is sometimes called a *white hole*. It turns out that regions II and II′ each have a second boundary, to which can be joined a fourth region I′. This is an exact replica of region I. Particles can pass out of region II′ into either of regions I and I′ or out of I or I′ into region II. However, there is no route by which a particle can pass from region I to region I′ or *vice versa*. Each of regions II and II′ has a real singularity at $r = 0$, which cannot be removed by any coordinate transformation. The one in region II is discussed below. The collection of four regions is called the *maximal extension* of the Schwarzschild solution. A description of the whole of this spacetime can be given by trading in both t and r for v and w, though there are other coordinate systems that do a better job. For a more detailed discussion of the Schwarzschild geometry, I must refer readers to more specialized books, such as Hawking and Ellis (1973) or Wald (1984).

So far in this section, our discussion has been purely mathematical: we have asked only about the geometrical structure implied by the Schwarzschild solution. We must now consider whether the curious phenomena associated with black and white holes can be brought about by known physical processes. Although the geometry described above represents an entire universe, this universe has to satisfy the assumptions that went into the Schwarzschild solution in the first place. This is obviously not true of our universe which, for example, contains more than one massive body. The most we can hope for in practice is that some fair-sized region in the neighbourhood of, say, a star is very similar to a corresponding region of the Schwarzschild spacetime.

The structure of a star is supported by its internal pressure and the outward flow of energy from nuclear reactions at its core. When its nuclear fuel is exhausted, the star collapses and, if it shrinks to a size equal to its own Schwarzschild radius, the conditions exist for the formation of a black hole. It appears, indeed, that once a mass is contained within its Schwarzschild radius, the gravitational attraction between its constituent parts cannot be counteracted by the outward pressure of any known force, and the mass is inevitably compressed to a single point—a singularity at $r = 0$. What becomes of this matter is not clear and readers should bear in mind that our whole discussion at this point ignores any quantum-mechanical considerations, which might profoundly affect the fate of the matter contained in a collapsing star.

From the point of view of the collapsing matter, the formation of the singularity occurs within a finite time although, as we shall see, the collapse appears to an external observer to take an infinite time. Theorems of Hawking and Penrose (discussed, for example, by Hawking and Ellis (1973)) show that

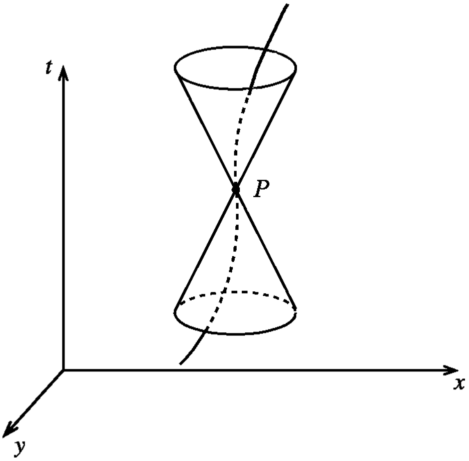


Figure 4.2. The light cone of a spacetime point P and a possible trajectory of a particle through P.

this phenomenon is rather general; for example, it does not depend on the exact spherical symmetry assumed by Schwarzschild. On the other hand, it seems likely that the geometry of the black holes formed by stellar collapse will usually not be of the Schwarzschild type, but rather will correspond to a *Kerr* solution, in which axial symmetry but not full spherical symmetry is assumed. This allows for the angular momentum possessed by a rotating star. Here, however, I shall consider only black holes of the simpler Schwarzschild type, which illustrate many of the same qualitative features. Notice that, prior to the stellar collapse, the exterior Schwarzschild solution we have considered is valid only outside the star, and therefore only for $r > r_S$. There is therefore no boundary at $r = r_S$ and $t = -\infty$ to which we might attach a region of type II$'$, and the question of forming a white hole does not arise. In fact there is not, to my knowledge, any physical process that is known to give rise to a white hole, and discussions of such objects are largely confined to the more speculative popular literature.

In Minkowski spacetime, the line element (2.6) implies that $|\mathrm{d}t/\mathrm{d}\tau| > |\mathrm{d}\boldsymbol{x}/\mathrm{d}\tau|$ along the path of any massive particle and that $|\mathrm{d}\boldsymbol{x}/\mathrm{d}t| = c$ for a light ray. As illustrated in figure 4.2, this implies that all possible light rays passing through a given point P lie on a cone, and that the path of a particle passing through P must be contained within this cone. This is expressed by saying that the path is *timelike* or, since the path is directed forwards in time, that it lies in the *forward light cone* of P. This is true both for freely falling particles and for those accelerated by some non-gravitational force. The familiar result of special relativity that no body can be accelerated past the speed of light is of course a direct consequence of this. Since any sufficiently small region of spacetime looks like Minkowski spacetime, the same is true of particle trajectories in every spacetime of physical interest.

The qualitative effect of black hole geometry on the paths of particles can be understood by plotting the paths of light rays and imagining particle trajectories to thread through the cones they produce. Using r and v to describe radial motion, we see from (4.49) that light rays, for which $\mathrm{d}\tau = 0$, satisfy

$$v = \text{constant} \tag{4.51}$$

or

$$\frac{\mathrm{d}r}{\mathrm{d}v} = \frac{1}{2}\left(1 - \frac{r_\mathrm{S}}{r}\right). \tag{4.52}$$

Readers may verify without difficulty that these curves are indeed null geodesics. In the case that v is constant, we find by differentiating (4.48) that $\mathrm{d}r/\mathrm{d}t = -c(1 - r_\mathrm{S}/r)$, so when r is large and t gives a measure of the time experienced by a stationary observer, we get $\mathrm{d}r/\mathrm{d}t \approx -c$. The set of curves corresponding to (4.51) therefore represent ingoing light rays. In figure 4.3, these curves are represented by diagonal lines from bottom right to top left.

Vertical lines are lines of constant r. The peculiarities of the geometry arise from the other set of light rays (4.52). One of these is the line $r = r_\mathrm{S}$, namely a ray that remains stationary at the Schwarzschild radius. Outside this radius, rays governed by (4.52) are outgoing; in fact, for these we find $\mathrm{d}r/\mathrm{d}t = c(1 - r_\mathrm{S}/r)$. Inside r_S, however, both sets of light rays fall inwards, terminating at the singularity at $r = 0$. Inside the Schwarzschild radius, therefore, all light rays and particles fall inwards. Events in region II are invisible to an outside observer, and the spherical surface at $r = r_\mathrm{S}$ (obtained by reinstating the angular coordinates) is called the *event horizon*.

The broken line in figure 4.3 represents the path of a particle falling from outside the event horizon. Suppose that it radiates light as it falls, so that a distant observer can follow its progress. It is apparent from the paths of the outgoing rays that this observer will have to wait an infinite time (measured for him by t) before receiving the signal emitted by the particle as it crosses the horizon. If light energy is radiated at a constant rate as measured by the proper time of the particle, then the finite amount of energy emitted in a short period just before the particle reaches the horizon is received by the observer over an infinite period of time. To him, therefore, the signal becomes ever fainter, and disappears entirely as the particle reaches the horizon. Also at this point, the interval between successive crests of a light wave becomes, for the observer, infinitely long so the light is infinitely redshifted.

Obviously, a black hole is, in itself, difficult to detect. On the other hand, if large amounts of matter are drawn in by the strong gravitational field that surrounds it, this matter may be expected to become very hot, giving rise to intense X- and γ radiation. This may happen, for example, in a binary star system, one of whose stars collapses to a black hole which can then accrete matter

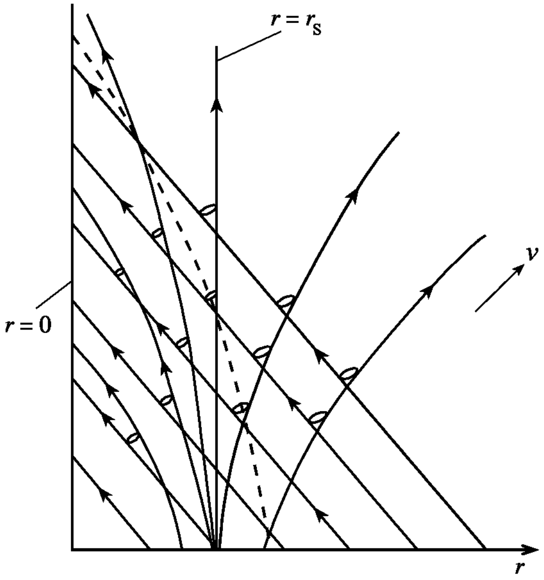


Figure 4.3. Trajectories of light rays (full lines) and an inward-falling particle (broken line) moving radially near a black hole.

from its companion. At the time of writing, there are numerous observed objects whose behaviour, in the view of many astronomers, provides strong circumstantial evidence of their containing black holes, though I know of no instances in which this identification is entirely unambiguous. Theory suggests that many stars which are bigger than a few solar masses will eventually collapse. In addition, large clusters of stars such as are found at the cores of galaxies appear to stand a good chance of coalescing to form very large black holes. In this connection, it is worthwhile to estimate the density of matter at the moment when an event horizon is formed. Suppose (although this is not strictly accurate) that the volume of this matter is just $4\pi r_S^3/3$, with r_S given in terms of the mass M by (4.29). Then its density can be estimated as

$$\rho \approx \frac{3c^6}{32\pi G^3 M_\odot^2}\left(\frac{M_\odot}{M}\right)^2 \approx \left(10^{16}\,\mathrm{g\,cm^{-3}}\right) \times \left(\frac{M_\odot}{M}\right)^2 \qquad (4.53)$$

where $M_\odot = 1.99 \times 10^{33}$ g is the mass of the Sun. If M is of the order of one solar mass, then this is an enormous density, which can be reached only at the core of a much larger object. On the other hand, if M is the combined mass of, say, 10^8 solar-mass stars (about 0.1% of the 10^{11} stars in an average galaxy) then this density is roughly that of water. All that is needed is that enough stars should accumulate in a ‘small’ region of space. There is evidence to suggest that this has in fact happened at the centre of our own galaxy.

Exercises

4.1. In a system of coordinates x^μ, let the coordinates of a point P be x_P^μ. If the connection coefficients are given by (2.50), show that, in a new coordinate system given by

$$x^{\mu'} = \delta_\mu^{\mu'}\left(x^\mu - x_P^\mu\right) + \tfrac{1}{2}\delta_\mu^{\mu'}\Gamma^\mu{}_{\nu\sigma}(x_P)\left(x^\nu - x_P^\nu\right)\left(x^\sigma - x_P^\sigma\right)$$

all first derivatives of the new components of the metric tensor field vanish at P.

4.2. The object of this exercise is to derive the field equations (4.17). Some of the results given in appendix A will be needed. The overall strategy is to make a small change in the metric, $g_{\mu\nu} \rightarrow g_{\mu\nu} + \delta g_{\mu\nu}$, and to require that the first-order change in the action (4.14) should vanish. The change in the gravitational part is

$$\delta S_{\text{grav}} = -\frac{1}{2c\kappa}\int \mathrm{d}^4x \Big[(2\Lambda + R)\delta\left((-g)^{1/2}\right) + (-g)^{1/2}\left(R_{\mu\nu}\delta g^{\mu\nu} + g^{\mu\nu}\delta R_{\mu\nu}\right)\Big].$$

(a) In the above expression, $\delta g^{\mu\nu}$ is the small change in the inverse metric $g^{\mu\nu}$. Let $\delta\bar{g}_{\mu\nu} = g_{\mu\alpha}g_{\nu\beta}\delta g^{\alpha\beta}$ be the quantity obtained by lowering its indices with the original metric. To first order in these small changes, show that $\delta\bar{g}_{\mu\nu} = -\delta g_{\mu\nu}$.
(b) Show that $\delta\left((-g)^{1/2}\right) = \frac{1}{2}(-g)^{1/2}g^{\mu\nu}\delta g_{\mu\nu}$.
(c) Show that the difference between two connections, such as $\Gamma(g)$ and $\Gamma(g+\delta g)$, is a tensor field.
(d) Show that

$$g^{\mu\nu}\delta R_{\mu\nu} = g^{\mu\nu}\left[\left(\delta\Gamma^\lambda{}_{\mu\nu}\right)_{;\lambda} - \left(\delta\Gamma^\lambda{}_{\mu\lambda}\right)_{;\nu}\right] = \left[g^{\mu\nu}\delta\Gamma^\lambda{}_{\mu\nu} - g^{\mu\lambda}\delta\Gamma^\nu{}_{\mu\nu}\right]_{;\lambda}.$$

Hence show that this term contributes to δS only a surface integral, which does not affect the field equations.
(e) Find the change in S_{matter} and complete the derivation of the field equations.

4.3. A radioactive material that emits photons of frequency ν is fixed to the roof of an elevator, which is initially at rest relative to a frame of reference S_{E} fixed in the Earth. At the instant that a photon is emitted vertically downwards, the elevator is released and begins to fall freely with acceleration g. After a short while, the photon hits a detector fixed to the floor of the elevator, having fallen a total distance h relative to S_{E}. Relative to S_{E}, how long did this take? According to the principle of equivalence, what frequency would the detector measure? Now suppose instead that the elevator has no floor, and what the photon actually hits is a detector fixed to the Earth's surface. What is the elevator's speed relative to S_{E} as the photon hits the detector? Since this is much smaller than c, use the non-relativistic Doppler formula to find the frequency ν' measured by this fixed detector. You should find that the fractional change in frequency is $(\nu' - \nu)/\nu = gh/c^2$, which comes to about 2.5×10^{-15} for a height of 22.6 m as used by Pound and Rebka. Using the approximation that h is much smaller than the radius of the Earth, verify that (4.33) gives the same result.

4.4. This exercise investigates the bending of light by the Sun, by considering the path of a light ray in the equatorial plane of the Schwarzschild spacetime, with coordinates (r, ϕ). First note that, in Euclidean space, the equation $r \sin\phi = r_0$ describes a straight line whose distance of closest approach to the origin is r_0. Along this line, $r \to \infty$ at $\phi = 0$ (corresponding to an approaching light ray) and at $\phi = \pi$ (corresponding to a departing light ray), while the point of closest approach is at $\phi = \pi/2$. This equation can be written as $u = \sin\phi/r_0$, where $u = 1/r$. In the Schwarzschild spacetime, let $u = 1/r$, where r is the *coordinate* that appears in (4.28) and let r_0 be the coordinate distance of closest approach.
(a) Recall that (4.40) and (4.41) are valid for a null geodesic, if $\mathrm{d}/\mathrm{d}\tau$ is replaced by differentiation with respect to a suitable parameter λ. Use these and (4.28) to derive the equation

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 + u^2(1 - r_S u) = (r_0 - r_S)/r_0^3.$$

(b) Treating $\epsilon = r_S/r_0$ as a small parameter, show that the solution to this equation for which $u = 0$ when $\phi = 0$ is approximately

$$r_0 u = \sin\phi + \tfrac{1}{2}\epsilon\left[(1 - \cos\phi)^2 - \sin\phi\right] + \mathrm{O}(\epsilon^2).$$

(c) Define the deflection angle α such that $u = 0$ when $\phi = \pi + \alpha$. Show that $\alpha = 2\epsilon + \mathrm{O}\left(\epsilon^2\right)$. Taking r_0 to be the solar radius 6.96×10^5 km (why is this allowed?), show that a light ray which just grazes the surface of the Sun is deflected by an angle of 1.75 seconds of arc.

4.5. Suppose that Mercury and the Earth could be frozen in their orbits at coordinate distances r_M and r_E in a direct line from the centre of the Sun. The distance between them can be found from (4.36) with r_S the Schwarzschild radius of the Sun. If the planets were separated by this distance in Euclidean space, what would be the round-trip time τ_{Euc} for a radar signal reflected from the surface of Mercury? In Schwarzschild spacetime, what is the coordinate time taken for the radar signal to complete the round trip? What is the corresponding time interval τ_{Sch} perceived by an observer on Earth? Taking r_M and r_E to be much larger than r_S, show that the general-relativistic time delay $\Delta\tau = \tau_{\mathrm{Sch}} - \tau_{\mathrm{Euc}}$ is given approximately by

$$\Delta\tau \approx \frac{r_S}{c}\left[\ln\left(\frac{r_E}{r_M}\right) + \left(\frac{r_M}{r_E}\right) - 1\right].$$

Estimate the magnitude of this effect by taking $r_M = 5.5 \times 10^7$ km and $r_E = 1.5 \times 10^8$ km.

4.6. A planet orbits a star whose Schwarzschild radius is r_S along a circular path with radial coordinate r. Verify that this is a geodesic of the Schwarzschild metric. Show that the coordinate time for one revolution is the same as the period of an orbit of radius r in the Newtonian theory. Show that a proper time interval experienced by the inhabitants of the planet is $(1 - 3r_S/2r)^{1/2}$ times the corresponding coordinate time interval.

4.7. Suppose that a photon of frequency ν can be considered as having kinetic energy $h\nu$ and the same gravitational potential energy as a particle of mass $h\nu/c^2$. Deduce the expression (4.33) for the frequency shift in a weak gravitational field. Do you think that such an interpretation could be rigorously justified? (Photons are discussed in chapter 5 and subsequent chapters.)

4.8. Show that a light ray can describe a circular orbit of coordinate radius $r = 3r_S/2$ around a black hole. How is this related to the result of exercise 4.6?

Chapter 5

Quantum Theory

Much of the remainder of this book will concern itself with those aspects of theoretical physics which seek to understand the nature of matter. Such understanding as we have has mainly been achieved by probing the structure of successively smaller constituents and, at least on the face of things, the regions of space and time we need to consider are far too small for spacetime curvature to be of any significance. Many of our considerations will therefore be restricted to Minkowski or, as in the present chapter, Galilean spacetime. Paradoxically, however, it seems that gravity and the structure of space and time may have a vital role to play in our understanding of matter on the very smallest scales, and we shall see something of the ways in which this comes about in later chapters.

In chapter 3, we studied some general theoretical aspects of classical or Newtonian mechanics which at the time seemed to provide a firm basis for understanding the properties and behaviour of material objects. As I hope readers are aware, it became apparent towards the end of the nineteenth century that a number of experimental observations could not be accommodated in this framework. As it turned out, a radical revision of both the mathematical and the conceptual foundations of mechanics is required to give an adequate account of these and subsequent observations, which arise most importantly in connection with atomic and subatomic phenomena. While the mathematical developments that constitute quantum mechanics have been outstandingly successful in describing all manner of observed properties of matter, it is fair to say that the conceptual basis of the theory is still somewhat obscure. I myself do not properly understand what it is that quantum theory tells us about the nature of the physical world, and by saying this I mean to imply that I do not think anybody else understands it either, though there are respectable scientists who write with confidence on the subject. This need not worry us unduly. There does exist a canon of generally accepted phrases which, if we do not examine them too critically, provide a reliable means of extracting from the mathematics well defined predictions for the outcome of any experiment we can perform (apart, that is, from the difficulty of solving the mathematical equations, which can be very great). I shall generally use these without comment, and readers must choose for themselves whether or not to accept them at face value.

This chapter deals with non-relativistic quantum mechanics, and I am going to assume that readers are already familiar with the more elementary aspects of the subject. The following section outlines the reasons why classical mechanics has proved inadequate and reviews the elementary ideas of wave mechanics. Although the chapter is essentially self-contained, readers who have not met this material before are urged to consult a textbook on quantum mechanics for a fuller account. The remaining sections develop the mathematical theory in somewhat more general terms, and this provides a point of departure for the quantum field theories to be studied in later chapters.

5.0 Wave Mechanics

The observations which led to the quantum theory are often summarized by the notion of *particle–wave duality*. Phenomena that might normally be regarded as wave motions turn out to have particle-like aspects, while particles behave in some respects like waves.

The phenomena in question are basically of three kinds. First, there is evidence that electromagnetic radiation, which for many purposes is described in terms of waves, behaves for other purposes like a stream of particles, called *photons*. (It is interesting to recall that Newton believed in a 'corpuscular' theory of light, propounded in his *Opticks*, but for reasons that have turned out to be quite erroneous.) In the photoelectric effect, for example, light incident on the surface of a metal causes electrons to be ejected. Contrary to what might have been expected, the energy of one of these electrons is found to be quite independent of the intensity of the radiation, although the number ejected per unit time does increase with the intensity. On the other hand, the energy of an electron increases with the frequency of the radiation. As Einstein was the first to realize, this can be understood if the radiation is considered to consist of photons, each carrying a definite amount of energy

$$E = h\nu \tag{5.1}$$

where ν is the frequency and $h = 6.6256 \times 10^{-34}\,\mathrm{J\,s}$ is Planck's constant. The energy of a single photon is transferred to a single electron, and the observed kinetic energy of the electron is this *quantum* of energy less a certain amount, the *work function*, required to release the electron from the metallic surface. Planck himself had been concerned with understanding the spectrum of black-body radiation, namely the way in which the energy radiated by a black object is distributed over frequencies. The analogous question of the distribution of molecular speeds in a gas could be well understood from a statistical analysis based on Newton's laws of motion, but this method failed when applied to electromagnetic waves. Planck discovered that, if the statistical analysis were to be modified by assuming that the energy carried by a wave of frequency ν

could only be some multiple of the quantum (5.1), then the correct spectrum could be obtained. Finally, the picture of radiation as a stream of particles is directly corroborated by the Compton effect, in which X rays scattered from electrons are found to undergo an increase in wavelength. According to electromagnetic theory which, as we have seen, is consistent with special relativity, a wave carrying energy E also carries a momentum $p = E/c$. If Compton scattering is viewed as a collision between a photon and an electron, then the change in wavelength is correctly found simply by requiring conservation of energy and momentum in each such collision. Since for electromagnetic radiation wavelength is related to frequency by $\lambda = c/\nu$, the momentum of a photon can be expressed as

$$p = h/\lambda \tag{5.2}$$

though as far as photons are concerned, this amounts merely to rewriting (5.1).

The second kind of evidence is that which shows that objects normally conceived of as particles have some wave-like properties. It was first suggested by de Broglie that the motion of a particle of energy E and momentum p might have associated with it a wave, whose frequency and wavelength would be given by (5.1) and (5.2). These would now be two independent equations, since the wave velocity would not, in general, be that of light. Celebrated experiments by Thomson and by Davisson and Germer showed that indeed electrons could be diffracted by a crystal lattice, just as light is by a diffraction grating, and confirmed the relation (5.2) between momentum and wavelength.

Lastly, there is the fact that atoms have definite ionization energies and radiate discrete rather than continuous spectra. This suggests that electrons in atoms occupy certain preferred orbits with definite allowed energies. If the electrons have waves associated with them, then the preferred states of motion can be envisaged as standing wave patterns, from which discrete energy levels arise in the same way as notes of a definite pitch from any musical instrument.

This talk of particle-wave duality may well strike readers as a leap in the dark. Indeed, it is undoubtedly the case that the elementary constituents of matter are neither particles nor waves, but rather entities of some other kind, for which our everyday experience provides no reliable analogy. Nevertheless, the de Broglie relations (5.1) and (5.2) point the way towards a quantitative theory that has become extraordinarily successful. I shall develop the essential points of this theory in more or less the traditional way, which should be made plausible, though it certainly is not justified in detail, by the experimental facts we have discussed.

Consider first a free particle, with energy E and 3-vector momentum $\boldsymbol{p}$. Classically, it would move in a straight line with constant velocity. With this motion, we must somehow associate a *wavefunction* $\Psi(\boldsymbol{x}, t)$ and since, according to (5.2) it must have a definite wavelength, the most reasonable guess for the nature of this wave is that it should be a plane wave. It turns out that wavefunctions must in general be complex, and a suitable guess is

$$\Psi(\boldsymbol{x}, t) = \exp\left[\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)\right]. \tag{5.3}$$

In terms of the angular frequency $\omega = 2\pi\nu$ and the wavevector $\boldsymbol{k}$, with $|\boldsymbol{k}| = 2\pi/\lambda$, we have $E = \hbar\omega$ and $\boldsymbol{p} = \hbar\boldsymbol{k}$, where $\hbar = h/2\pi$. We see at once that, since this wave exists everywhere in space, there is nothing to tell us where the particle is. The accepted interpretation is that, in general, the quantity

$$P(\boldsymbol{x}, t)\mathrm{d}^3x = |\Psi(\boldsymbol{x}, t)|^2\mathrm{d}^3x \tag{5.4}$$

is the probability of finding the particle, at time t, in an infinitesimal region d^3x surrounding the point $\boldsymbol{x}$. Alternatively, we can refer to $P(\boldsymbol{x}, t)$ itself as the *probability density* for finding the particle in the neighbourhood of $\boldsymbol{x}$. This means that the integral over all space of P should be 1. Therefore, (5.3) is not quite satisfactory as it stands, since it gives the value 1 for P itself. One method of modifying (5.3) is to suppose that the particle is confined to some large region of space and to divide the right-hand side of (5.3) by the square root of this volume.

More generally, if we wish to predict the result of a measurement of some quantity, say A, given that the state of motion of our system is described by a known wavefunction Ψ, it may well be that Ψ does not yield any particular value for A. In that case, we must be content with calculating probabilities for the measurement to yield various possible values of A. How such probabilities are obtained will be discussed in the next section. Clearly, however, we must have some means of extracting from the wavefunction whatever information it contains about the quantity A. To this end, we associate with every physical quantity a *differential operator*, which can act on any wavefunction. For the cases of energy and momentum, these are taken to be

$$\text{energy operator: } \mathrm{i}\hbar\frac{\partial}{\partial t} \tag{5.5}$$

$$\text{momentum operator: } -\mathrm{i}\hbar\boldsymbol{\nabla}. \tag{5.6}$$

Obviously, acting with these on the wavefunction (5.3) is equivalent to multiplying the wavefunction by E or $\boldsymbol{p}$ respectively. Other wavefunctions, corresponding to states in which the particle does not have a uniquely defined energy or momentum, can be written as superpositions of waves of the form (5.3) by Fourier transformation. If we act with the above operators on such a wavefunction, we obtain a new wavefunction in which each component of the superposition has been multiplied by its own energy or momentum. In a manner that will become clear below, we can compare the new wavefunction with the old one, or with plane waves, and by making these comparisons we obtain all the information that quantum mechanics allows us to have about the energy or momentum of the particle in the given state of motion.

To find out how the state of motion of a system evolves with time, we can, in simple cases at least, make use of the fact that its energy can be expressed in terms of other quantities. For example, if we have a single particle of mass m moving in a potential $V(\boldsymbol{x})$, then its energy is $E = (\boldsymbol{p}^2/2m) + V(\boldsymbol{x})$. By substituting the operators (5.5) and (5.6) into this equation, and allowing each side to act on the wavefunction, we obtain *Schrödinger's equation*

$$\mathrm{i}\hbar \frac{\partial}{\partial t}\Psi(\boldsymbol{x},t) = \left(-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{x})\right)\Psi(\boldsymbol{x},t). \tag{5.7}$$

With these preliminary ideas in mind, we can proceed to develop the mathematical theory in detail. One of our main concerns will be to show how equations (5.5) and (5.6), which we obtained more or less by guesswork, can be justified at a deeper level in terms of the symmetries that we studied in chapter 3.

5.1 The Hilbert Space of State Vectors

In order to develop the theory of classical mechanics, we had first to decide how a unique state of a physical system could be specified, and this question must now be reconsidered. We have already seen that, if a quantum-mechanical particle has a definite momentum, then it cannot also have a definite position. More generally, there will be *maximal sets* of observable quantities, say $\{A, B, C, \ldots\}$, such that every quantity in the set can, at the same time, have a definite value, while any other quantity either is forbidden to have a definite value at the same time, or has a value that is determined by the values of $A, B, C, \ldots$. For a single free particle whose only properties are position and momentum, $\{\boldsymbol{x}\}$ and $\{\boldsymbol{p}\}$ are examples of such maximal sets. The energy $E = \boldsymbol{p}^2/2m$ does not count, because it can be expressed in terms of $\boldsymbol{p}$. We shall say that a system is specified to be in a *pure quantum state* when all the values $\{a, b, c, \ldots\}$ of quantities belonging to some maximal set have been given. The criterion for deciding which sets of observables actually *are* maximal sets will emerge later on.

The first crucial assumption we made in chapter 3 for classical mechanics was that every instantaneous state could be specified in terms only of the positions and velocities of all the particles of the system. We now need a corresponding assumption for quantum mechanics, which again can ultimately be justified only by the fact that it leads to successful predictions about experimental observations. It consists in the following enigmatic statement:

> all possible instantaneous states of the system can be represented by vectors in a Hilbert space.

The mathematical definition of a Hilbert space is given in appendix A, and the properties of these spaces are discussed in many mathematical textbooks (see, for example, Simmons (1963)). For many purposes in physics, however, it is enough to think of state vectors as a straightforward generalization of ordinary Euclidean 3-vectors and I shall follow this line of thought, ignoring a number of subtleties that must be taken into account in a fully rigorous treatment. The main generalizations are:

(i) The Hilbert space can have any number of dimensions, and we usually need an infinite number to accommodate all possible states.

(ii) A 3-vector can be multiplied by any positive real number α, the effect being to multiply its length by α, leaving its direction unchanged, or by a negative number which reverses the direction. A state vector may be multiplied by any complex number.

(iii) We denote a state vector by $|\Psi\rangle$, the Ψ being simply a label for identification. The scalar product $\boldsymbol{u} \cdot \boldsymbol{v}$ of two 3-vectors generalizes to a complex number $\langle\Phi|\Psi\rangle$, which has the property

$$\langle\Psi|\Phi\rangle = \langle\Phi|\Psi\rangle^*. \tag{5.8}$$

In a sense, we might understand (ii) as saying that the length of a vector is allowed to be complex. However, the length of a vector $|\Psi\rangle$ as defined by mathematicians is $\sqrt{\langle\Psi|\Psi\rangle}$, which is a real number.

Suppose for the moment that each observable quantity in the maximal set $\{A, B, C, \ldots\}$ can assume only a discrete set of values. The state in which these values are $a, b, c, \ldots$ will be represented by a vector $|a, b, c, \ldots\rangle$ normalized so that $\langle a, b, c, \ldots|a, b, c, \ldots\rangle = 1$. Each of the vectors obtained by multiplying this one by any non-zero complex number corresponds to the same physical state, and the set of all such vectors is called a *ray*. Thus, each physical state corresponds to a ray or, in other words, a *direction* in the Hilbert space. The relationship between the quantum state of a system and physical measurements performed on it is the subject of the following basic postulate of the theory. Suppose the actual state is represented by a vector $|\Psi\rangle$, normalized so that $\langle\Psi|\Psi\rangle = 1$, and a measurement is made of all the quantities in some maximal set. Then the probability of obtaining the set of results $\{a, b, c, \ldots\}$ is

$$P(a, b, c, \ldots|\Psi) = |\langle a, b, c, \ldots|\Psi\rangle|^2. \tag{5.9}$$

Clearly, the goal of quantum-mechanical calculations will be to find these scalar products, though we do not yet know how to set about this. Readers who have studied chapter 2 will appreciate that the existence of scalar products implies that the Hilbert space possesses a structure analogous to a metric, and that this gives a unique correspondence between a vector $|\Psi\rangle$ and a one-form $\langle\Psi|$ which is the other half of the scalar product symbol, sometimes called a *dual vector*. (Readers who have studied section 3.7 should note that this use of the term 'dual' is not quite the same as the one used there.) In less formal language, it is generally convenient to think of two Hilbert spaces, which carry exactly the same information, differently packaged. One is composed of vectors $|\Psi\rangle$ and the other of dual vectors $\langle\Psi|$. Exercise 5.1 uses the algebra of complex matrices to show how this works in concrete terms. A whimsical terminology due to P A M Dirac calls $|\ \rangle$ a 'ket' and $\langle\ |$ a 'bra', so that the scalar product becomes a bra(c)ket. I shall express the one-to-one correspondence between bra and ket vectors by writing

$$\langle\Psi| = |\Psi\rangle^\dagger \qquad \text{and} \qquad |\Psi\rangle = \langle\Psi|^\dagger \tag{5.10}$$

although the † symbol is more properly reserved for use with operators as described below. The property (5.8) of the scalar product implies that, if α is a complex number, then

$$(\alpha|\Psi\rangle)^\dagger = \alpha^*\langle\Psi| \qquad \text{and} \qquad (\alpha\langle\Psi|)^\dagger = \alpha^*|\Psi\rangle. \tag{5.11}$$

If $|\Psi\rangle$ is the state $|a', b', c', \ldots\rangle$, where a and a' are two possible values of A, and so on, then the probability in (5.9) must be equal to 1 if the two sets of values are the same and zero otherwise. This implies that two state vectors associated with the same maximal set of observables are *orthonormal*, which means

$$\langle a, b, c, \ldots|a', b', c', \ldots\rangle = \delta_{aa'}\delta_{bb'}\delta_{cc'}\cdots. \tag{5.12}$$

On the other hand, the total probability of getting *some* set of values from the measurement is found by summing (5.9) over all possible values of $a, b, c, \ldots$ and must be equal to 1. This will be true if every state vector can be expressed as a sum of the form

$$|\Psi\rangle = \sum_{a,b,c,\ldots} \psi_{abc\ldots}|a, b, c, \ldots\rangle. \tag{5.13}$$

If $|\Psi\rangle$ is normalized, the complex coefficients in this expression satisfy

$$\langle\Psi|\Psi\rangle = \sum_{a,b,c,\ldots} |\psi_{abc\ldots}|^2 = 1 \tag{5.14}$$

and readers may easily verify, using (5.12), that the sum of probabilities (5.9) is indeed 1. If $|\Psi\rangle$ is not normalized, then the right-hand side of (5.9) must be divided by $\langle\Psi|\Psi\rangle$.

The fact that every state vector can be expressed in the form (5.13) means that the set of vectors $|a, b, c, \ldots\rangle$ associated with a maximal set of observables forms an orthonormal basis for the Hilbert space. Choosing a new set of basis vectors, corresponding to a different maximal set of observables, is like rotating the coordinate axes in Euclidean geometry.

If one of the observables, say A, can assume a continuous range of values, then $\delta_{aa'}$ in (5.12) must be replaced by the Dirac function $\delta(a-a')$ and the sums in (5.13) and (5.14) by integrals. As far as A is concerned, the probability (5.9) then becomes a probability density, in the sense discussed in the last section. Consider, for example, a single particle, and choose the maximal set to be $\{\boldsymbol{x}\}$. Although a state vector is not the same thing as a wavefunction, a given state of motion can be represented either by a state vector $|\Psi\rangle$ or by a wavefunction $\psi(\boldsymbol{x})$. In fact, if $|\boldsymbol{x}\rangle$ represents the state in which the particle has exactly the position $\boldsymbol{x}$, then the wavefunction is simply the coefficient of $|\boldsymbol{x}\rangle$ in the expansion

$$|\Psi\rangle = \int \mathrm{d}^3x\, \psi(\boldsymbol{x})|\boldsymbol{x}\rangle. \tag{5.15}$$

Since the orthonormality condition is now $\langle\boldsymbol{x}|\boldsymbol{x}'\rangle = \delta^3(\boldsymbol{x} - \boldsymbol{x}')$, we get

$$\psi(\boldsymbol{x}) = \langle\boldsymbol{x}|\Psi\rangle \tag{5.16}$$

and for the probability density we find

$$P(\boldsymbol{x}|\Psi) = |\psi(\boldsymbol{x})|^2. \tag{5.17}$$

Apart from the fact that we are not yet dealing with time evolution, this agrees exactly with (5.4).

5.2 Operators and Observable Quantities

Suppose we have a rule that enables us to associate with any given vector $|\Psi\rangle$ another vector $|\Psi'\rangle$. We say that an *operator* $\hat{O}$ acts on $|\Psi\rangle$ to produce $|\Psi'\rangle$:

$$|\Psi'\rangle = \hat{O}|\Psi\rangle. \tag{5.18}$$

I shall usually use the circumflex to indicate operators. The rule that defines an operator may be specified in various ways, and sometimes rather indirect means are necessary since it is impractical to consider each vector of the Hilbert space individually. The simplest operator of all is the *identity* operator $\hat{I}$, which leaves every vector unchanged. Almost all the operators used in quantum theory are *linear*. This means that, for any two vectors $|\Phi\rangle$ and $|\Psi\rangle$ and any two complex numbers α and β, we have

$$\hat{O}\left(\alpha|\Phi\rangle + \beta|\Psi\rangle\right) = \alpha\hat{O}|\Phi\rangle + \beta\hat{O}|\Psi\rangle. \tag{5.19}$$

All operators in this book are linear unless otherwise stated.

Observable quantities can be represented by operators in the following way. Let A belong to a maximal set $\{A, B, C, \ldots\}$. If the state of the system is one of the corresponding basis vectors $|a, b, c, \ldots\rangle$ then A has the definite value a, and we define the action of an operator $\hat{A}$ on each basis vector to be that of multiplying it by a:

$$\hat{A}|a, b, c, \ldots\rangle = a|a, b, c, \ldots\rangle. \tag{5.20}$$

An equation of this form, in which the action of an operator is just to multiply the vector by a number, is called an *eigenvalue equation* We say that $|a, b, c, \ldots\rangle$ is an *eigenvector* of $\hat{A}$ with *eigenvalue* a. Since any vector can be expanded as in (5.13), this tells us how $\hat{A}$ acts on every vector. The probability $P(a|\Psi)$ of getting the result a from a measurement of A, irrespective of the values of any other quantities, is found by summing (5.9) over all the values of $b, c, \ldots$. Readers should be able to verify that the *expectation value* $\langle A\rangle$, which means the average value of A obtained from many measurements, is

$$\langle A\rangle = \sum_a aP(a|\Psi) = \langle\Psi|\hat{A}|\Psi\rangle. \tag{5.21}$$

The expression on the right-hand side means the scalar product of $\langle\Psi|$ with the vector $\hat{A}|\Psi\rangle$.

In view of the symmetrical appearance of expressions like this, it is useful to define the action of operators on bra vectors also. The new bra vector $\langle\Phi|\hat{A}$ is defined by requiring that, for any $\langle\Phi|$ and any $|\Psi\rangle$, the expression $\langle\Phi|\hat{A}|\Psi\rangle$ has the same value, whether we regard it as the scalar product of $\langle\Phi|\hat{A}$ and $|\Psi\rangle$ or of $\langle\Phi|$ and $\hat{A}|\Psi\rangle$. For the reason discussed in exercise 5.1 (which readers may like to study before proceeding), this quantity is called a *matrix element* of $\hat{A}$. There is a second method by which an operator may be used to obtain a new bra vector. If $\langle\Psi|$ is the bra whose corresponding ket is $|\Psi\rangle$, we can first form the new ket vector $\hat{A}|\Psi\rangle$ and then find its corresponding bra. The new bras formed by these two methods are not necessarily the same. We can describe the second method in terms of the action of an operator $\hat{A}^{\dagger}$, which is called the *adjoint* or the *Hermitian conjugate* of $\hat{A}$:

$$\left(\hat{A}|\Psi\rangle\right)^{\dagger} = \langle\Psi|\hat{A}^{\dagger}. \tag{5.22}$$

Using (5.8), we find that for any two vectors

$$\langle\Psi|\hat{A}|\Phi\rangle = \langle\Phi|\hat{A}^{\dagger}|\Psi\rangle^{*}. \tag{5.23}$$

An operator which equals its own adjoint

$$\hat{A}^{\dagger} = \hat{A} \tag{5.24}$$

is called *self-adjoint* or *Hermitian*. Strictly speaking, these two terms have slightly different meanings, but the distinction will not concern us.

In (5.23), let us take $\hat{A}$ to be Hermitian, $|\Phi\rangle$ to be an eigenvector of $\hat{A}$ with eigenvalue a_1 and $|\Psi\rangle$ an eigenvector with eigenvalue a_2. We find

$$\left(a_1 - a_2^{*}\right)\langle\Psi|\Phi\rangle = 0. \tag{5.25}$$

In the case that $|\Phi\rangle = |\Psi\rangle$, we have $a_2 = a_1$, so we see that the eigenvalues of an Hermitian operator are real. On the other hand, if the two eigenvalues are different, then the two eigenvectors must be orthogonal (which means $\langle\Psi|\Phi\rangle = 0$). These two properties are just what we need if $\hat{A}$ is to represent a measurable quantity, since its eigenvalues are possible results of measurements and therefore real numbers, and we want its eigenvectors to satisfy (5.12). We therefore assume that all observable quantities are represented by Hermitian operators.

The sum of two operators is defined so as to be consistent with the addition of two vectors. That is, to act with $(\hat{A} + \hat{B})$ on a vector $|\Psi\rangle$, we first act with $\hat{A}$ and $\hat{B}$ separately and then add the resulting vectors: $(\hat{A}+\hat{B})|\Psi\rangle = \hat{A}|\Psi\rangle + \hat{B}|\Psi\rangle$.

The product $\hat{A}\hat{B}$ of two operators represents the combined effect of acting on a ket vector with $\hat{B}$ and then acting on the resulting vector with $\hat{A}$: $\hat{A}\hat{B}|\Psi\rangle = \hat{A}(\hat{B}|\Psi\rangle)$. The product $\hat{B}\hat{A}$, in which $\hat{A}$ acts before $\hat{B}$, does not necessarily have the same effect. The difference between these two operators is another operator, called the *commutator* of $\hat{A}$ and $\hat{B}$ and written as

$$[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}. \tag{5.26}$$

In practice, most of the information we have about operators derives from *commutation relations*, which express commutators in terms of other operators. This is largely because of the role played by commutators in the symmetry operations discussed in the next section. We can use the definition of a commutator to express the criterion for building the maximal sets of observables from which our discussion started. If A and B belong to the same set, then acting on one of the associated basis vectors $|a, b, c, \ldots\rangle$ with $\hat{A}$ and $\hat{B}$ in either order gives the same result, namely multiplying it by ab. Since this is true for every basis vector, the result of acting with $\hat{A}$ and $\hat{B}$ in either order on any vector is the same. Therefore, their commutator is zero and they are said to *commute*. ('Zero' here means the operator that acts on any vector to give the vector whose length is zero.) Thus, a maximal set of observables is such that all the corresponding operators commute with each other, and no other independent operator commutes with all of them (except $\hat{I}$, which commutes with everything).

We shall often need to consider operators that are functions of other operators. To illustrate what is involved, consider the expression $\hat{A} = \exp(\alpha\hat{B})$. Since we know how to multiply operators, we can make sense of this by using the power series expansion

$$\hat{A} = \hat{I} + \alpha\hat{B} + \tfrac{1}{2!}\alpha^2\hat{B}^2 + \cdots . \tag{5.27}$$

For some purposes, we can treat this function as if $\hat{B}$ were a number. For example, the inverse operator $\hat{A}^{-1}$ is defined by $\hat{A}^{-1}\hat{A} = \hat{A}\hat{A}^{-1} = \hat{I}$. It is equal to $\exp(-\alpha\hat{B})$, as may readily be verified by multiplying the two series together. On the other hand, readers may verify in the same way that the product $\exp(\hat{B})\exp(\hat{C})$ is not equal to $\exp(\hat{B} + \hat{C})$ unless $\hat{B}$ and $\hat{C}$ commute. Obviously, functions of operators must be handled with care. A power series is often the best way of resolving doubts as to whether a particular manipulation is permissible. By using (5.23), we find that the adjoint of $\hat{A} = \exp(\alpha\hat{B})$ is $\hat{A}^\dagger = \exp(\alpha^*\hat{B}^\dagger)$. If $\alpha = \mathrm{i}$ and $\hat{B}$ is Hermitian, this implies that

$$\hat{A}^\dagger = \hat{A}^{-1} \tag{5.28}$$

in which case $\hat{A}$ is said to be *unitary*.

5.3 Spacetime Translations and the Properties of Operators

In order to make use of the formalism we have developed so far, we obviously need information about the specific properties of operators that represent particular physical quantities. The only way to acquire this information is to make informed guesses and see whether they lead to a successful theory. Our only guide in this enterprise is classical mechanics, and I propose to make the required guesses as plausible as possible by drawing analogies with the discussions of chapter 3. We begin with time translations.

There are several different ways of describing the evolution in time of the state of a system. The most obvious, which we consider first, is called the *Schrödinger picture*. Each vector in the Hilbert space is associated with a possible instantaneous state of the system, so we denote by $|\Psi(t)\rangle$ its state at time t. If we suppose that the initial state $|\Psi(0)\rangle$ at time $t = 0$ is known, then the relation between these two states can be described by a *time evolution operator* $\hat{U}(t)$:

$$|\Psi(t)\rangle = \hat{U}(t)|\Psi(0)\rangle. \tag{5.29}$$

In order to preserve the probabilistic interpretation of $|\Psi(t)\rangle$ in a systematic way, we require its normalization to remain constant:

$$\langle\Psi(t)|\Psi(t)\rangle = \langle\Psi(0)|\hat{U}^{\dagger}(t)\hat{U}(t)|\Psi(0)\rangle = \langle\Psi(0)|\Psi(0)\rangle. \tag{5.30}$$

Evidently, $\hat{U}(t)$ must be a unitary operator with $\hat{U}(0) = \hat{I}$ and, according to our discussion at the end of the last section, it can be written as

$$\hat{U}(t) = \exp(-\mathrm{i}\hat{\mathcal{H}}t) \tag{5.31}$$

where $\hat{\mathcal{H}}$ is an Hermitian operator. If we assume that $\hat{\mathcal{H}}$ is independent of time, insert (5.31) into (5.29) and differentiate, we get

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}|\Psi(t)\rangle = \hat{\mathcal{H}}|\Psi(t)\rangle \tag{5.32}$$

which has the same form as the Liouville equation (3.22) for the evolution of the state in classical mechanics. Now $\mathcal{H}$ in (3.22) was a differential operator constructed from the Hamiltonian function, which is usually the same as the total energy. The quantum-mechanical operator $\hat{\mathcal{H}}$ is Hermitian, and therefore suitable for representing an observable quantity. A reasonable guess, therefore, is that $\hat{\mathcal{H}}$ is proportional to the quantum-mechanical Hamiltonian or total energy operator $\hat{H}$. Since the argument of the exponential in (5.31) must be dimensionless, our guess is

$$\hat{\mathcal{H}} = \hbar^{-1}\hat{H} \tag{5.33}$$

where $\hbar$ is a fundamental constant with the dimensions of energy × time. The value of this constant must eventually be determined experimentally, and it turns out, of course, to be none other than Planck's constant divided by 2π.

A different view of time evolution, called the *Heisenberg picture*, comes about when we realize that $|\Psi(t)\rangle$ is not itself an observable quantity. The expectation value of an observable quantity at time t can be written without reference to $|\Psi(t)\rangle$ as

$$\langle\Psi(t)|\hat{A}|\Psi(t)\rangle = \langle\Psi|\hat{A}(t)|\Psi\rangle \tag{5.34}$$

where $|\Psi\rangle$ means $|\Psi(0)\rangle$ and

$$\hat{A}(t) = \hat{U}^{\dagger}(t)\hat{A}\hat{U}(t) = \exp(\mathrm{i}\hat{\mathcal{H}}t)\hat{A}\exp(-\mathrm{i}\hat{\mathcal{H}}t). \tag{5.35}$$

The two operators $\hat{A}$ and $\hat{A}(t)$ have their analogues in classical mechanics where, as we have seen, a function $A(\{q\}, \{p\})$ defines the meaning of a given dynamical quantity in terms of coordinates and momenta, whereas $A(t) = A(\{q(t)\}, \{p(t)\})$ gives the value of this quantity when we substitute for $\{q\}$ and $\{p\}$ the actual solutions of the equations of motion. These solutions depend on the initial values $\{q(0)\}$ and $\{p(0)\}$ and substituting definite numerical values for these yields a definite function for $A(t)$ corresponding to an entire history of the system as it evolves from the chosen initial state. The quantum-mechanical analogue of inserting these initial conditions is to form the expectation value $\langle\Psi|\hat{A}(t)|\Psi\rangle$. In this sense, $|\Psi\rangle$ represents an entire history of the quantum-mechanical system, from which we extract time-dependent information using the Heisenberg-picture operators $\hat{A}(t)$. We can easily derive an equation of motion for $\hat{A}(t)$, analogous to (3.17), by differentiating (5.35). Since $\hat{\mathcal{H}}$ obviously commutes with $\hat{U}$ and $\hat{U}^\dagger$, these can be differentiated as if $\hat{\mathcal{H}}$ were a number. But, since $\hat{\mathcal{H}}$ does not necessarily commute with $\hat{A}$, we must be careful about the order of operators in the result, which is

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{A}(t) = -\mathrm{i}[\hat{A}(t), \hat{\mathcal{H}}] = -\frac{\mathrm{i}}{\hbar}[\hat{A}(t), \hat{H}]. \tag{5.36}$$

An immediate consequence of this is that any quantity whose associated operator commutes with $\hat{H}$ is conserved. In particular, $\hat{H}$ commutes with itself and is conserved. The assumption that went into this result was that $\hat{\mathcal{H}}$, and therefore the quantum-mechanical law of motion, did not depend explicitly on time. In view of our discussion in §3.2, we would expect conservation of energy to be an automatic consequence of this assumption, which reinforces our interpretation of $\hat{H}$ as representing the total energy.

In chapter 3, we constructed from the total momentum an operator $\mathcal{P}$ (equation(3.24)) which generates translations in space just as $\mathcal{H}$ does in time. From considerations similar to those above, we can ascertain the properties of the corresponding quantum-mechanical operator. Comparing (5.36) with (3.17), we observe a correspondence of the form

$$[\hat{A}, \hat{B}] = i\hbar\widehat{\{A, B\}}_{\mathrm{P}} \tag{5.37}$$

where the right-hand side means that we first evaluate the Poisson bracket in terms of classical coordinates and momenta and then substitute the corresponding quantum-mechanical operators. If this correspondence were generally true, the definition (3.18) of the Poisson bracket would imply, in particular, the *canonical commutation relations*

$$[\hat{x}_\alpha, \hat{p}_\beta] = \mathrm{i}\hbar\delta_{\alpha\beta} \tag{5.38}$$

$$[\hat{x}_\alpha, \hat{x}_\beta] = 0 \tag{5.39}$$

$$[\hat{p}_\alpha, \hat{p}_\beta] = 0 \tag{5.40}$$

where α and β label the Cartesian components of particle positions and momenta. On the right-hand sides of these equations, and in similar contexts, we understand a complex number to mean the operator that multiplies a vector by this number. I shall shortly give arguments that make the commutation relations (5.38)-(5.40) fairly plausible. These commutation relations comprise the whole of our knowledge about momentum and position operators, and indeed the entire theory of quantum mechanics rests on (5.36) and (5.38)-(5.40). It should be emphasized, though, that the correspondence (5.37) does not necessarily hold in general. If we know the commutator $[\hat{x}_\alpha, \hat{p}_\beta]$, then we can work out the commutator of any two operators $\hat{A}$ and $\hat{B}$ constructed from the coordinates and momenta. Quite often, the result will be found to agree with (5.37), but this is not necessarily so.

To obtain the commutation relations (5.38)-(5.40), recall that in classical mechanics the generator of space translations $\mathcal{P} = \mathrm{i}\{\boldsymbol{P}, \quad\}_{\mathrm{P}}$ is related to the total momentum $\boldsymbol{P}$ in the same way that the generator of time translations $\mathcal{H} = \mathrm{i}\{H, \quad\}_{\mathrm{P}}$ is related to the Hamiltonian. Having guessed that the quantum-mechanical generator of time translations is to be identified through (5.33), we now make the consistent assumption that

$$\hat{\mathcal{P}} = \hbar^{-1}\hat{\boldsymbol{P}} = \hbar^{-1}\sum_i \hat{\boldsymbol{p}}_i \tag{5.41}$$

where $\hat{\boldsymbol{p}}_i$ is the linear momentum operator for the ith particle. From this generator, we can construct a space translation operator $\exp(-\mathrm{i}\boldsymbol{a}\cdot\hat{\mathcal{P}})$, analogous to the time evolution operator (5.31), which displaces the system through a vector $\boldsymbol{a}$. Again, the argument of this exponential must be dimensionless, so it is important to note that the dimensions of $\hbar$ can be expressed as momentum × distance. To simplify matters, I shall deal just with a single particle, so that $\hat{\mathcal{P}} = \hbar^{-1}\hat{\boldsymbol{p}}$, but readers should not find it hard to convince themselves that the argument extends to a system of many particles also.

For the moment, I propose to accept (5.39), which asserts that the components of the particle's position commute with each other, on the intuitive grounds that all three of these components ought to be simultaneously measurable. The assertion of (5.40), that the three momentum components also commute, might seem justifiable on the same grounds but, for reasons that I shall discuss later, we need to consider this more carefully. If the components of $\hat{\boldsymbol{p}}$ commute with each other, then

$$\exp(-\mathrm{i}\boldsymbol{a}\cdot\hat{\mathcal{P}})\exp(-\mathrm{i}\boldsymbol{b}\cdot\hat{\mathcal{P}}) = \exp[-\mathrm{i}(\boldsymbol{a}+\boldsymbol{b})\cdot\hat{\mathcal{P}}]. \tag{5.42}$$

This means that a translation through a vector $\boldsymbol{b}$ followed by a translation through a vector $\boldsymbol{a}$ is equivalent to a single translation through the vector $\boldsymbol{a}+\boldsymbol{b}$, as it ought to be. The fundamental reason for requiring the momentum components to commute with each other is to preserve this property of space translations.

Now consider an operator $\hat{A} = A(\hat{\boldsymbol{x}})$ which is a function just of the position operator $\hat{\boldsymbol{x}}$. By analogy with (5.35), the effect of a space translation on this operator is

$$A(\hat{\boldsymbol{x}}+\boldsymbol{a}) = \exp(\mathrm{i}\boldsymbol{a}\cdot\hat{\mathcal{P}})A(\hat{\boldsymbol{x}})\exp(-\mathrm{i}\boldsymbol{a}\cdot\hat{\mathcal{P}}). \tag{5.43}$$

For the particular case $A(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}$, this becomes

$$\hat{\boldsymbol{x}} + \boldsymbol{a} = \exp(\mathrm{i}\boldsymbol{a} \cdot \hat{\boldsymbol{\mathcal{P}}})\hat{\boldsymbol{x}} \exp(-\mathrm{i}\boldsymbol{a} \cdot \hat{\boldsymbol{\mathcal{P}}}) \tag{5.44}$$

and if $|\boldsymbol{x}\rangle$ and $|\boldsymbol{x} + \boldsymbol{a}\rangle$ are eigenvectors of $\hat{\boldsymbol{x}}$, with

$$\hat{\boldsymbol{x}}|\boldsymbol{x}\rangle = \boldsymbol{x}|\boldsymbol{x}\rangle \qquad \text{and} \qquad \hat{\boldsymbol{x}}|\boldsymbol{x} + \boldsymbol{a}\rangle = (\boldsymbol{x} + \boldsymbol{a})|\boldsymbol{x} + \boldsymbol{a}\rangle$$

then we can deduce from (5.44) the action of the translation operator on $|\boldsymbol{x}\rangle$, namely

$$\exp(-\mathrm{i}\boldsymbol{a} \cdot \hat{\boldsymbol{\mathcal{P}}})|\boldsymbol{x}\rangle = |\boldsymbol{x} + \boldsymbol{a}\rangle. \tag{5.45}$$

Let us expand (5.44) in powers of $\boldsymbol{a}$ and use our guess that $\hat{\boldsymbol{\mathcal{P}}} = \hbar^{-1}\hat{\boldsymbol{p}}$. The terms linear in $\boldsymbol{a}$ on each side must be equal, so we find

$$a_\alpha = -\frac{\mathrm{i}}{\hbar}\sum_\beta a_\beta[\hat{x}_\alpha, \hat{p}_\beta] \tag{5.46}$$

and this implies that the commutator is given by (5.38). Using this relation, exercise 5.3 shows that (5.44) and (5.43) are true to all orders in $\boldsymbol{a}$. Accepting that the commutation relation (5.38) is correct, we see that $[\hat{H}, \hat{\boldsymbol{p}}] = 0$ if and only if $\hat{H}$ is independent of $\hat{\boldsymbol{x}}$; that is, if and only if the system is translationally invariant. In that case, the equation of motion (5.36) with $\hat{A} = \hat{\boldsymbol{p}}$ shows that the momentum is conserved.

By now, we can see a general pattern emerging. In classical mechanics, we can identify quantities that are conserved for a system that is invariant under the various spacetime symmetry transformations discussed in chapter 3. In quantum mechanics, the operators that represent these quantities are to be identified (up to a factor of $\hbar$) as the generators of the corresponding transformations, and this determines their commutation properties. It is instructive to see how these ideas apply to rotations, which we have not yet considered. According to exercise 3.1, the conserved quantity associated with rotations is the angular momentum, whose components (with the notation $\boldsymbol{x} = (x, y, z)$) are

$$\hat{J}_x = \hat{y}\hat{p}_z - \hat{z}\hat{p}_y \qquad \hat{J}_y = \hat{z}\hat{p}_x - \hat{x}\hat{p}_z \qquad \hat{J}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x. \tag{5.47}$$

Quantum-mechanically, the rotation generators found in that exercise are indeed given by $\hat{\boldsymbol{\mathcal{J}}} = \hat{\boldsymbol{x}} \times \hat{\boldsymbol{\mathcal{P}}} = \hbar^{-1}\hat{\boldsymbol{x}} \times \hat{\boldsymbol{p}} = \hbar^{-1}\hat{\boldsymbol{J}}$. Using the commutation relations we have found for $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{p}}$, it is straightforward to work out the commutators of the angular momentum components

$$[\hat{J}_x, \hat{J}_y] = \mathrm{i}\hbar\hat{J}_z \qquad [\hat{J}_y, \hat{J}_z] = \mathrm{i}\hbar\hat{J}_x \qquad [\hat{J}_z, \hat{J}_x] = \mathrm{i}\hbar\hat{J}_y. \tag{5.48}$$

For the classical angular momentum, on the other hand, we can work out the corresponding Poisson brackets and verify that (5.37) is true. Evidently, the three components of angular momentum do not commute with each other. This reflects the fact that two consecutive rotations about different axes do not in general produce the same result if their order is reversed. Had we been content, earlier on, to accept (5.40) on the grounds that the three components of momentum ought to be simultaneously measurable, then the same reasoning ought to have applied to angular momentum, and this would have led to inconsistent results. It would seem, then, that this argument also stands on dangerous ground when applied to the position operators. Now that we have understood the whole scheme, perhaps the best that can be said is that the correspondence (5.37) between Poisson brackets and commutators, when applied to Cartesian coordinates and to the generators of spacetime symmetry transformations, provides a mathematically consistent, and reasonably plausible basis for further investigation.

5.4 Quantization of a Classical System

Until we have some experience of quantum-mechanical systems, the only sensible way we have of specifying such a system is to model it upon a classical one. Given the formal correspondences we have seen to exist between classical and quantum mechanics, it is not difficult to give a prescription for 'quantizing' a classical system. It is called the *canonical quantization* scheme. Usually, the classical system can be specified by giving its Lagrangian as a function of generalized coordinates $\{q_i\}$ and their velocities. The momentum $p_i = \partial L/\partial \dot{q}_i$ conjugate to each coordinate can be found and the velocities eliminated in favour of the momenta. The Hamiltonian can then be found as in §3.3. Finally, the quantum-mechanical system can be defined by replacing the coordinates and momenta with the corresponding operators and requiring these operators to satisfy the commutation relations

$$[\hat{q}_i, \hat{p}_j] = \mathrm{i}\hbar\delta_{ij}. \tag{5.49}$$

These relations apply to Schrödinger-picture operators or to Heisenberg-picture operators *at the same time*. The commutator $[\hat{q}_i(t), \hat{p}_j(t')]$ is equal to $\mathrm{i}\hbar\delta_{ij}$ if and only if $t = t'$, as readers are invited to prove. If $t \neq t'$, its value depends on how the system has evolved between these two times and is different for systems with different Hamiltonians. In most cases, no simple expression can be found for it.

When implementing this procedure, one may encounter ambiguities of various kinds, and satisfactory methods of dealing with these must be sought. It is possible, for example, that different choices of the generalized coordinates, which would yield equivalent descriptions of a classical system, may produce inequivalent results when the commutation relations (5.49) are imposed. For systems of non-relativistic particles, at least, the safe course seems to be to use Cartesian coordinates. When the classical Hamiltonian contains products of variables whose corresponding operators do not commute, the quantum Hamiltonian is not unambiguously prescribed. A possible course is to replace, say, $\hat{A}\hat{B}$ with the symmetrized product $\frac{1}{2}(\hat{A}\hat{B} + \hat{B}\hat{A})$, but other solutions may be appropriate in specific cases. A further difficulty arises if the time derivative of some coordinate does not appear in the Lagrangian. The momentum conjugate to this coordinate is identically zero and (5.49) obviously cannot hold. Ordinarily, this does not happen when the classical Lagrangian describes a system of particles, because the kinetic energy term involves all the velocity components. It does happen, however, when we try to extend the formalism to treat the electromagnetic field as a quantum system, and for systems that are subject to constraints of various sorts (see Lawrie and Epp (1996) for a simple example).

A point worth noting is that velocities do not, in general, have a well defined meaning in quantum mechanics. We have seen that, if a particle has a definite momentum, its position is completely undetermined. To assign it a velocity would require two exact measurements of its position, separated by an infinitesimal time interval, which does not make good quantum-mechanical sense, even as an idealized limiting process. The momenta that appear in (5.49) are always the canonically defined ones. In the presence of electromagnetic forces, for example, they correspond to classical quantities of the kind shown in (3.59) (though we have not yet given a proper account of the quantum mechanics of relativistic particles) rather than to just $m\dot{\boldsymbol{x}}$.

Although the formulation of quantum mechanics in terms of state vectors and operators acting on them is more general than wave mechanics, the solution of specific problems is often most conveniently achieved in terms of wavefunctions. Let us therefore see how the algebra of operators acting on state vectors can be reinterpreted in terms of differential operators on wavefunctions. The wavefunction corresponding to a state vector $|\Psi\rangle$ is given by (5.16). The wavefunction corresponding to $\hat{\boldsymbol{x}}|\Psi\rangle$ is

$$\langle \boldsymbol{x}|\hat{\boldsymbol{x}}|\Psi\rangle = \boldsymbol{x}\langle \boldsymbol{x}|\Psi\rangle = \boldsymbol{x}\psi(\boldsymbol{x}) \tag{5.50}$$

and so the action of the Schrödinger-picture position operator corresponds to multiplication of the wavefunction by the coordinate. Similarly, using (5.45), with $\hat{\boldsymbol{\mathcal{P}}} = \hat{\boldsymbol{p}}/\hbar$, we can write

$$\exp(\boldsymbol{a}\cdot\nabla)\psi(\boldsymbol{x}) = \psi(\boldsymbol{x}+\boldsymbol{a}) = \langle \boldsymbol{x}|\exp(\mathrm{i}\boldsymbol{a}\cdot\hat{\boldsymbol{p}}/\hbar)|\Psi\rangle. \tag{5.51}$$

As in (3.23), the exponential of the gradient operator represents a Taylor series. Clearly, the action of $\hat{\boldsymbol{p}}$ on $|\Psi\rangle$ corresponds to that of $-\mathrm{i}\hbar\nabla$ on the wavefunction. Readers should be able to satisfy themselves that, given any operator which can be expressed as a function $A(\hat{\boldsymbol{x}},\hat{\boldsymbol{p}})$, the wavefunction corresponding to the vector $A(\hat{\boldsymbol{x}},\hat{\boldsymbol{p}})|\Psi\rangle$ is $A(\boldsymbol{x},-\mathrm{i}\hbar\nabla)\psi(\boldsymbol{x})$. In particular, if A is the Hamiltonian for a particle moving in the potential $V(\boldsymbol{x})$, we see from (5.32) that the time-dependent wavefunction $\psi(\boldsymbol{x},t) = \langle \boldsymbol{x}|\Psi(t)\rangle$ obeys Schrödinger's equation (5.7). To complete the correspondence between state vectors and wavefunctions, we note first that the operators $\boldsymbol{x}$ and $-\mathrm{i}\hbar\nabla$ satisfy the same commutation relations (5.38) as $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{p}}$ and second, as readers may show, that any matrix element may be expressed as

$$\langle\Phi|A(\hat{\boldsymbol{x}})|\Psi\rangle = \int \mathrm{d}^3x\,\phi^*(\boldsymbol{x})A(\boldsymbol{x},-\mathrm{i}\hbar\nabla)\psi(\boldsymbol{x}). \tag{5.52}$$

The extension of these considerations to systems containing more than one particle, with wavefunctions $\psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots) = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots |\Psi\rangle$, should be obvious.

5.5 An Example: The One-Dimensional Harmonic Oscillator

The harmonic oscillator provides a standard illustration of the mathematical ideas we have developed. It also serves to introduce the idea of raising and lowering operators, which are of fundamental importance for second quantization and field theory, which we study in the following chapter. The classical system from which we start consists of a single particle of mass m, moving in one dimension in the potential $V(x) = \frac{1}{2}m\omega^2 x^2$, and the classical trajectories are sinusoidal oscillations of angular frequency ω. The Lagrangian is

$$L = \tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}m\omega^2 x^2. \tag{5.53}$$

The momentum conjugate to x is $p = m\dot{x}$ and the Hamiltonian is

$$H = \frac{1}{2m}p^2 + \frac{1}{2}m\omega^2 x^2. \tag{5.54}$$

None of the above-mentioned difficulties occurs here, so we are free to impose the commutation relation $[\hat{x}, \hat{p}\,] = \mathrm{i}\hbar$.

We developed the mathematics of state vectors and operators by assuming that a Hilbert space describing all possible states of motion of our system was given, and enquiring about the properties of operators that act on it. Now, however, we see that the practical problem of theoretical physics is the reverse: our physical principles supply us with operators having definite properties, and we have to construct a Hilbert space by finding the states of motion that are permitted by these properties. This problem will be solved if we can find a set of basis vectors and if we know how any operator acts on each basis vector. A set of basis vectors will be associated with some maximal set of observables, and the most useful sets are $\{x\}$, $\{p\}$ and $\{H\}$. The description in terms of a particular set of basis vectors is called a *representation*, and the representations associated with the above maximal sets are called, logically enough, the coordinate, momentum and energy representations.

We shall first construct the basis vectors for the energy representation. These are eigenvectors of the Hamiltonian, labelled by an integer n, with eigenvalues ϵ_n:

$$\hat{H}|n\rangle = \epsilon_n |n\rangle. \tag{5.55}$$

They are of particular interest because they are *stationary states*. Time-dependent vectors of the form $\exp(-\mathrm{i}\epsilon_n t)|n\rangle$ are solutions of (5.32), and the expectation value in such a state of any operator that is defined in a time-independent manner is constant. If, for example, the oscillator is regarded as a model for the vibrations of a diatomic molecule, then the observed spectral lines arise from transitions between these states, caused by external forces that are not included in our description. The fact that the allowed energy levels have discrete values rather than a continuous range is at present a matter of assumption, but will be verified in due course. It is advantageous to exchange the position and momentum for two new operators $\hat{a}$ and $\hat{a}^\dagger$, defined by

$$\hat{a} = \left(\frac{\omega m}{2\hbar}\right)^{1/2}\left[\hat{x} + \left(\frac{1}{\omega m}\right)\mathrm{i}\hat{p}\right] \tag{5.56}$$

$$\hat{a}^\dagger = \left(\frac{\omega m}{2\hbar}\right)^{1/2}\left[\hat{x} - \left(\frac{1}{\omega m}\right)\mathrm{i}\hat{p}\right] \tag{5.57}$$

in terms of which the Hamiltonian can be written as

$$\hat{H} = \left(\hat{a}^\dagger\hat{a} + \tfrac{1}{2}\right)\hbar\omega. \tag{5.58}$$

Using the commutation relation for $\hat{x}$ and $\hat{p}$, we find that these operators satisfy

$$[\hat{a}, \hat{a}^\dagger] = 1 \tag{5.59}$$

$$[\hat{a}, \hat{H}] = \hbar\omega\hat{a} \tag{5.60}$$

$$[\hat{a}^\dagger, \hat{H}] = -\hbar\omega\hat{a}^\dagger. \tag{5.61}$$

From (5.60), it is easy to show that if $|n\rangle$ is an energy eigenvector with energy ϵ_n, then $\hat{a}|n\rangle$ is an eigenvector with energy $(\epsilon_n - \hbar\omega)$. In fact, we can calculate

$$\hat{H}(\hat{a}|n\rangle) = \hat{H}\hat{a}|n\rangle = (\hat{a}\hat{H} - \hbar\omega\hat{a})|n\rangle = (\epsilon_n - \hbar\omega)(\hat{a}|n\rangle). \tag{5.62}$$

Similarly, (5.61) implies that $\hat{a}^\dagger|n\rangle$ is an eigenvector with energy $(\epsilon_n + \hbar\omega)$. For this reason, $\hat{a}$ and $\hat{a}^\dagger$ are called energy lowering and raising operators.

Written in terms of $\hat{x}$ and $\hat{p}$, the Hamiltonian is a sum of squares of Hermitian operators. Therefore, none of its eigenvalues can be negative, and there must be a *ground state* of minimum energy, which we denote by $|0\rangle$. Since $\hat{a}|0\rangle$ cannot be a state with lower energy, the only way to satisfy (5.60) when it acts on $|0\rangle$ is to have $\hat{a}|0\rangle = 0$. Then, acting on $|0\rangle$ with the Hamiltonian (5.58) shows that $\epsilon_0 = \frac{1}{2}\hbar\omega$. By acting n times on $|0\rangle$ with $\hat{a}^\dagger$, we generate an infinite series of energy eigenvectors with energies

$$\epsilon_n = \left(n + \tfrac{1}{2}\right)\hbar\omega. \tag{5.63}$$

Furthermore, there cannot be any states with energies between these values. If there were, then by acting enough times with $\hat{a}$, we could generate a state with energy between 0 and $\hbar\omega$, but not equal to $\frac{1}{2}\hbar\omega$. Acting once more with $\hat{a}$ would have to produce zero, by the same argument as before. But we already know that a state with this property has an energy of exactly $\frac{1}{2}\hbar\omega$, which is a contradiction. Thus, the states $|n\rangle$, with energy eigenvalues given by (5.63), constitute the complete set of basis vectors for the energy representation. We require these basis vectors to be normalized so that $\langle n|n'\rangle = \delta_{nn'}$. I leave it as an exercise for readers to establish (by induction) that they are given by

$$|n\rangle = (n!)^{-1/2}(\hat{a}^\dagger)^n|0\rangle \tag{5.64}$$

and that

$$\hat{a}^\dagger|n\rangle = (n+1)^{1/2}|n+1\rangle \qquad \text{and} \qquad \hat{a}|n\rangle = n^{1/2}|n-1\rangle. \tag{5.65}$$

This is, essentially, the solution to our problem. Any observable property of the oscillator can be expressed in terms of $\hat{x}$ and $\hat{p}$, and it is a trivial matter to express these in terms of $\hat{a}$ and $\hat{a}^\dagger$ by solving (5.56) and (5.57). Any state vector can be expressed as a linear combination of the basis vectors $|n\rangle$, and so (5.65) tells us how any operator acts on any vector. A particularly useful operator is $\hat{n} = \hat{a}^\dagger\hat{a}$, which has the property

$$\hat{n}|n\rangle = \hat{a}^\dagger\hat{a}|n\rangle = n|n\rangle. \tag{5.66}$$

It is called the *number operator*, because it counts the number of *quanta* $\hbar\omega$ of energy in the state.

These results may be translated into the coordinate representation by finding the wavefunctions $\psi_n(x)$ of the energy eigenstates. The two sets of basis vectors are related by

$$|n\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, \psi_n(x)|x\rangle \qquad \text{and} \qquad |x\rangle = \sum_{n=0}^{\infty} \psi_n^*(x)|n\rangle. \tag{5.67}$$

To find the wavefunctions, we rewrite the raising and lowering operators in terms of x and $-\mathrm{i}\hbar\partial/\partial x$. The ground-state wavefunction $\psi_0(x)$ is the solution of the equation $a(x, -\mathrm{i}\hbar\partial/\partial x)\psi_0(x) = 0$, and the others are found by applying the raising operator to it. The result may be written as

$$\psi_n(x) = N_n \exp\left(\frac{\omega m x^2}{2\hbar}\right)\left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^n \exp\left(-\frac{\omega m x^2}{\hbar}\right) \tag{5.68}$$

where the normalizing factor

$$N_n = \left[n!\left(\frac{\pi\hbar}{\omega m}\right)^{1/2}\left(\frac{2\omega m}{\hbar}\right)^n\right]^{-1/2}$$

ensures that

$$\int_{-\infty}^{\infty} |\psi_n(x)|^2\mathrm{d}x = 1. \tag{5.69}$$

A further translation into the momentum representation is simply a matter of Fourier transformation. It can easily be verified that the relations

$$|x\rangle = (2\pi\hbar)^{-1/2}\int \mathrm{d}p\, \exp(-\mathrm{i}px/\hbar)|p\rangle \tag{5.70}$$

$$|p\rangle = (2\pi\hbar)^{-1/2}\int \mathrm{d}x\, \exp(\mathrm{i}px/\hbar)|x\rangle \tag{5.71}$$

are uniquely determined by (5.45) and the orthonormality requirements $\langle x|x'\rangle = \delta(x - x')$ and $\langle p|p'\rangle = \delta(p - p')$. Consequently, the energy eigenvectors may be expressed as

$$|n\rangle = \int \mathrm{d}p\, \pi_n(p)|p\rangle \tag{5.72}$$

where the *momentum-space wavefunction* is

$$\pi_n(p) = (2\pi\hbar)^{-1/2} \int \mathrm{d}x\ \exp(-\mathrm{i}px/\hbar)\psi_n(x). \tag{5.73}$$

Obviously, this method of solving the problem works only for the particular case of the harmonic oscillator. For single particles in other potentials, the most practical method of constructing the Hilbert space is to use the coordinate representation. The eigenvalue equation (5.55) becomes the *time-independent Schrödinger equation*

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{x})\right]\psi_\epsilon(\boldsymbol{x}) = \epsilon\psi_\epsilon(\boldsymbol{x}). \tag{5.74}$$

In the case of the harmonic oscillator, the boundary conditions on the solutions of this equation are that the wavefunction must vanish sufficiently fast as $|x| \to \infty$ for the integral in (5.69) to converge to a finite value, which can be normalized to 1. This is possible only when ϵ has one of the values in (5.63), so it is these boundary conditions that lead to the energy of the oscillator being quantized in a set of discrete levels. In all these states, the probability density (5.17) vanishes rapidly when $|x|$ becomes sufficiently large. In this sense, the particle is constrained by the parabolic potential to remain close to the origin, and the states are known as *bound states*.

In almost every physical problem, the potential approaches a finite value, which might as well be zero, at infinity. The Coulomb potential seen by the electron in a hydrogen atom is an archetypical example. If the potential possesses a well, then there may be bound states of negative energy, in which the particle is most probably to be found in the well. The spectrum of bound-state energy levels is always discrete. In positive-energy states, however, the particle can escape to infinity, where the wave function becomes similar to (5.3). These are called *scattering states*. The energies of scattering states form a continuous spectrum, because different boundary conditions apply to them. The exact nature of these boundary conditions is slightly complicated, because the wavefunctions cannot be made to satisfy (5.69) or its three-dimensional equivalent. In fact, if the particle is not bound by the potential, the usefulness of the energy eigenfunctions associated with the potential is limited, and a different description is appropriate. I shall return briefly to this question in chapter 9 and in appendix D.

The use of wave functions to solve both bound state and scattering problems is of the utmost importance in many areas of physics. The practical techniques available are described in any respectable textbook on quantum mechanics, but they are not part of the subject matter of this book, and I must ask readers to look elsewhere for further details.

Exercises

5.1. The object of this exercise is to show that manipulation of state vectors and operators is entirely analogous to the algebra of complex matrices and is in fact identical in the case of a Hilbert space of finite dimension. Readers are invited to satisfy themselves of this, and to gain some further insight, by considering the various assertions made below. Little or no detailed working may be needed. Let $|\psi\rangle$ stand for the column matrix $(\psi_1, \ldots, \psi_N)^{\mathrm{T}}$, where the ψ_i are complex numbers and $^{\mathrm{T}}$ denotes the transpose. An orthonormal basis is given by the vectors $|i\rangle$, where $|1\rangle = (1, 0, 0, \ldots, 0)^{\mathrm{T}}$, $|2\rangle = (0, 1, 0, \ldots, 0)^{\mathrm{T}}$ and so on.

(a) Any column matrix $|\psi\rangle$ can be expressed as a linear combination of the basis vectors $|i\rangle$, with coefficients ψ_i.

(b) If α is any complex number, then $\alpha|\psi\rangle = (\alpha\psi_1, \ldots, \alpha\psi_N)^{\mathrm{T}}$.

(c) If $\langle\psi|$ is the row matrix $(\psi_1^*, \ldots, \psi_N^*)$, and $\langle\psi|\phi\rangle$ is the usual matrix product, then (5.8) and (5.11) are true.

(d) Multiplication by any $N \times N$ square matrix $\hat{A}$ provides a rule for converting any column matrix into another column matrix.

(e) Any square matrix can be multiplied on the left by a row matrix, and the elements of $\hat{A}$ are $\hat{A}_{ij} = \langle i|\hat{A}|j\rangle$.

(f) If the elements of $\hat{A}^\dagger$ are $(\hat{A}^\dagger)_{ij} = \hat{A}^*_{ji}$, then (5.22) and (5.23) are true.

(g) If $\hat{A}|i\rangle = a_i|i\rangle$ for each basis vector, then $\hat{A}$ is a diagonal matrix with diagonal elements a_i.

(h) If $\hat{A}$ is a diagonal matrix, $\hat{B}$ is a square matrix such that $[\hat{A}, \hat{B}] = 0$, and $a_i \neq a_j$, then $\hat{B}_{ij} = 0$.

(i) If $\{\hat{A}, \hat{B}, \hat{C}, \ldots\}$ is a maximal set of operators (square matrices) in the sense discussed following (5.26), and the basis vectors $|i\rangle$ are their simultaneous eigenvectors, then $\hat{A}$, $\hat{B}$, $\hat{C}$, … are all diagonal and, for any pair of indices i and j, there is at least one member of the set whose ith and jth eigenvalues are not equal.

(j) If $\hat{A}$ is a diagonal matrix with diagonal elements a_i, then $f(\hat{A})$ is the diagonal matrix whose elements are $f(a_i)$.

5.2. For any set of operators $\hat{A}$, $\hat{B}$, $\hat{C}$, …, show that $(\hat{A}\hat{B}\hat{C}\cdots)^\dagger = \cdots\hat{C}^\dagger\hat{B}^\dagger\hat{A}^\dagger$ and $(\hat{A}\hat{B}\hat{C}\cdots)^{-1} = \cdots\hat{C}^{-1}\hat{B}^{-1}\hat{A}^{-1}$.

5.3. For a single coordinate and its conjugate momentum, use the canonical commutator (5.38) to show by induction that $\hat{x}\hat{p}^n = \hat{p}^n\hat{x} + n\mathrm{i}\hbar\hat{p}^{n-1}$ and $\hat{p}\hat{x}^n = \hat{x}^n\hat{p} - n\mathrm{i}\hbar\hat{x}^{n-1}$. Hence show, for any function f that has a Taylor expansion, that $\hat{x}f(\hat{p}) = f(\hat{p})\hat{x} + \mathrm{i}\hbar f'(\hat{p})$ and $\hat{p}f(\hat{x}) = f(\hat{x})\hat{p} - \mathrm{i}\hbar f'(\hat{x})$. Use these results to verify (5.44) and (5.43). For a system of several particles, whose potential energy depends only on the relative coordinates of pairs of particles, show that the total momentum is conserved.

5.4. The symbol $|\Psi\rangle\langle\Psi|$ represents a *projection operator*, which acts on any ket vector $|\Phi\rangle$ to produce the new ket vector $(\langle\Psi|\Phi\rangle)|\Psi\rangle$ and analogously on any bra vector. Show that the probability (5.9) is the expectation value of a projection operator. If $|a, b, c, \ldots\rangle$ are a complete set of basis vectors, show that their projection operators form a *resolution of the identity*, which means that

$$\sum_{a,b,c,\ldots} |a, b, c, \ldots\rangle\langle a, b, c, \ldots| = \hat{I}.$$

Show that the operator $\hat{A}$, for which $\hat{A}|a, b, c, \ldots\rangle = a|a, b, c, \ldots\rangle$, can be expressed as

$$\hat{A} = \sum_{a,b,c,\ldots} |a, b, c, \ldots\rangle a\langle a, b, c, \ldots|.$$

How can this be generalized to represent an operator that is not diagonal in this representation?

5.5. If $f'(x)$ denotes the derivative $\mathrm{d}f(x)/\mathrm{d}x$ when x is an ordinary number, show that $\mathrm{d}f(\alpha\hat{A})/\mathrm{d}\alpha = \hat{A}f'(\alpha\hat{A})$.

5.6. Let $|i\rangle$ and $|\alpha\rangle$ be two sets of orthonormal basis vectors such that

$$|i\rangle = \sum_{\alpha} u_{i\alpha}|\alpha\rangle.$$

Show that the complex coefficients $u_{i\alpha}$ are the components of a unitary matrix.

5.7. Let $\hat{A}$ and $\hat{B}$ be two operators such that the commutator $\hat{C} = [\hat{A}, \hat{B}]$ commutes with both $\hat{A}$ and $\hat{B}$, and let $:\cdots:$ denote an ordering of operators such that $\hat{A}$s always stand to the left of $\hat{B}$s. So, for example,

$$:(\hat{A} + \hat{B})^n: = \sum_{m=0}^{n} \binom{n}{m} \hat{A}^m \hat{B}^{n-m}$$

where $\binom{n}{m}$ is the binomial coefficient.

(a) Show by induction that

$$(\hat{A} + \hat{B})^{n+1} = \hat{A}(\hat{A} + \hat{B})^n + (\hat{A} + \hat{B})^n\hat{B} - n\hat{C}(\hat{A} + \hat{B})^{n-1}.$$

(b) Show that

$$(\hat{A} + \hat{B})^n = \sum_{m=0}^{[n/2]} \alpha_{nm}\hat{C}^m :(\hat{A} + \hat{B})^{n-2m}:$$

where $[n/2]$ equals $n/2$ if n is even or $(n-1)/2$ if n is odd, and the expansion coefficients satisfy the recursion relation

$$\alpha_{n+1,m+1} = \alpha_{n,m+1} - n\alpha_{n-1,m}.$$

(c) Verify that this recursion relation is solved by

$$\alpha_{nm} = \left(-\frac{1}{2}\right)^m \frac{n!}{(n-2m)!m!}$$

and hence derive the *Baker-Campbell-Hausdorff formula*

$$\exp(\hat{A}+\hat{B}) = \exp(\hat{A})\exp(\hat{B})\exp(-\hat{C}/2).$$

(d) Show that $\exp(\hat{A})\exp(\hat{B}) = \exp(\hat{B})\exp(\hat{A})\exp(\hat{C})$.

Chapter 6

Second Quantization and Quantum Field Theory

Up to a point, the quantum theory developed in chapter 5 was quite general. However, the systems we had in mind were non-relativistic ones consisting of a fixed number of particles. In this chapter, we extend the theory to deal with systems in which the number of particles can change. There are several reasons for wanting to do this. The most obvious is that we need a method of describing high-energy scattering and decay processes in which particles can be created and destroyed. A second is that, when we try to make quantum theory consistent with special relativity, we encounter difficulties (discussed in chapter 7) that can be resolved only in this more general setting. The final reason is that, even for systems of non-relativistic particles, the mathematics rapidly becomes intractable as the number of particles increases. A useful device for dealing with large systems is, roughly speaking, to imagine adding an extra particle, which serves as a theoretical probe of the state of the system. To put the matter another way, the method of *second quantization* developed in this chapter provides a means of dealing with the entire system by considering only a few particles at a time.

The term ‘second quantization’ is an unfortunate one, insofar as it suggests a theory which is ‘twice as quantum-mechanical’ as the one we started with. This is emphatically not the case: all we are doing is developing a convenient mathematical technique for dealing with the original theory. The origin of the term will become clear, but briefly it is this. Addition or subtraction of particles to or from the system is represented by creation and annihilation operators, which are closely analogous to the raising and lowering operators of the harmonic oscillator. From these we can construct *field operators* which, in the absence of interactions, satisfy the same Schrödinger equation as single-particle wavefunctions. By turning a wavefunction, which is acted on by operators representing physical quantities, into an operator which itself acts on state vectors, we might appear to be adding a further layer of quantumness, but readers who follow the development carefully will realize that this is not a good description of what is actually taking place.

6.1 The Occupation-Number Representation

Consider a system containing a fixed number N of identical particles. For the moment, we shall assume that they do not interact with each other. Some, though not all, states of the system can be specified by giving the state of motion of each particle. I shall label a complete set of single-particle states by the symbol k. Quite often, it will be convenient to take these single-particle states to be momentum eigenstates, in which case k will represent the value of the momentum. Other sets of states, such as the Bloch states that describe the motion of electrons in a crystal lattice, may be more useful in particular circumstances. Also, if the particles have spin, then the spin state of the particle is included in k. (Readers who are not familiar with spin will find a brief discussion in appendix B, and I shall discuss its relativistic origin in chapter 7; those who are not familiar with Bloch states may like to consult a book on solid state physics, but need not do so for the purposes of this book.) Thus, if we choose to specify the momentum (k_x, k_y, k_z) and spin s of an electron, then k is a shorthand for this set of four numbers.

Using these single-particle states, we can choose a set of basis vectors for the whole system of the form $|k_1, k_2, \ldots, k_N\rangle$, where the nth label in the list refers to the nth particle. Because quantum-mechanical particles do not follow definite trajectories, it is impossible in principle to distinguish two identical particles. Therefore, the two vectors $|k_1, k_2, \ldots\rangle$ and $|k_2, k_1, \ldots\rangle$ must be taken as referring to the same physical state and can differ only by a phase factor. That is, $|k_2, k_1, \ldots\rangle = \alpha|k_1, k_2, \ldots\rangle$, where α is a complex number of unit magnitude. On interchanging the particles a second time, we get back to the original vector, so $\alpha^2 = 1$. The same is true, of course, for any pair of particles. The state is said to be *symmetric* if $\alpha = 1$ or *antisymmetric* if $\alpha = -1$. It is found that particles with integral spin can exist only in symmetric states. They are said to obey *Bose–Einstein statistics* and are called *bosons*. Particles with half-odd-integer spin exist only in antisymmetric states. They obey *Fermi–Dirac statistics* and are called *fermions*. The only known explanation for this state of affairs (the *spin-statistics theorem*) comes from relativistic field theories and will be touched on in chapter 7.

For the moment, we deal only with bosons. The order of k labels in a basis vector is immaterial: the same set of labels in any order identifies the same vector. It is a simple matter to allow for variable numbers of particles to be present. We simply include in the Hilbert space state vectors with arbitrary numbers of k labels. The orthonormality condition for these vectors is a bit cumbersome to write down correctly. I shall exhibit an expression for it, and then explain its meaning. The expression is

$$\begin{aligned}&\langle k_1, k_2, \ldots, k_N | k'_1, k'_2, \ldots, k'_{N'}\rangle \\ &\qquad = C\delta_{NN'} \sum_P \delta(k_1 - k'_{P(1)})\delta(k_2 - k'_{P(2)}) \cdots \delta(k_N - k'_{P(N)}).\end{aligned} \tag{6.1}$$

We want this scalar product to be zero unless the two vectors represent the same physical state. They must first of all have the same number of particles, which accounts for $\delta_{NN'}$. Then, we need delta function constraints to ensure that each vector represents the same set of single-particle states. Each delta function in (6.1) stands for a product of deltas, one for each variable represented by k: a Kronecker symbol for a discrete variable and a Dirac function for a continuous one. If we list the k labels of a given vector in a different order, we still have the same vector. Therefore, we must arrange matters so that the constraints will be satisfied if any permutation of the labels $k'_1, \ldots, k'_N$ matches the set $k_1, \ldots, k_N$. In (6.1), the set of numbers $P(1), \ldots, P(N)$ is a permutation of $1, \ldots, N$, and we achieve the desired effect by summing over all permutations. If, say, n of the k_i are equal, then $n!$ of the terms in this sum will simultaneously be satisfied and, to get the correct normalization, we divide by $n!$. If there are several sets of equal k_i, then we divide by the $n!$ for each set, and this normalization factor is denoted by C.

If at least one of the variables represented by k is continuous, it will be extremely rare for two of the k_i to have exactly the same value, and C is almost always equal to 1. In mathematical terms, the Dirac delta function makes good sense only when it appears inside an integral and, for readers who understand such matters, 'almost always' means 'except on a set of zero measure'. It often happens that all the variables in k have only a discrete set of values. For example, if the particles are confined to a cubical box of side L, then each momentum component can have only the values $2\pi\hbar n/L$, where n is an integer. In that case, it is possible to use a different notation in which $k_1, k_2, \ldots$ are the allowed values of k, rather than the k associated with different particles. The basis vectors can then be denoted by $|n_1, n_2, \ldots\rangle$, where n_i is the number of particles in the state k_i. This is called the *occupation-number representation*, the n_i being the occupation numbers of single-particle states. The orthonormality condition can be written much more straightforwardly as

$$\langle n_1, n_2, \ldots | n'_1, n'_2, \ldots\rangle = \delta_{n_1 n'_1}\delta_{n_2 n'_2}\cdots. \tag{6.2}$$

At this point, it is interesting to note the greater generality of the formulation of quantum theory in terms of state vectors as opposed to wavefunctions. In the Schrödinger picture, the time-dependent state of the system is some linear combination of basis vectors

$$|\Psi(t)\rangle = \sum_{n_1, n_2, \ldots} \Psi_{n_1 n_2 \ldots}(t)|n_1, n_2, \ldots\rangle. \tag{6.3}$$

In a quite natural way, this represents in general a superposition of states in which the system has different numbers of particles. If the system does contain a fixed number N of particles, then only those coefficients for which the occupation numbers add to N will be non-zero. If the Hamiltonian does not allow for processes in which particles are created or destroyed, then this number will be conserved. If such processes are possible, then even if we start with a definite number of particles, the number remaining after some period of time will be uncertain, and the superposition will contain states with different numbers of remaining particles. This situation cannot be represented by a wavefunction, which necessarily has a definite number of arguments.

It is now possible to introduce creation and annihilation operators, which convert a given basis vector into one with an extra particle or one with a particle missing. In the occupation-number representation, the process is precisely analogous to changing the number of quanta of energy in the state of an harmonic oscillator. For each single-particle state k, we define operators $\hat{a}(k)$ and $\hat{a}^\dagger(k)$ by

$$\hat{a}(k_i)|n_1, n_2, \ldots, n_i, \ldots\rangle = n_i^{1/2}|n_1, n_2, \ldots, (n_i - 1), \ldots\rangle \tag{6.4}$$

$$\hat{a}^\dagger(k_i)|n_1, n_2, \ldots, n_i, \ldots\rangle = (n_i + 1)^{1/2}|n_1, n_2, \ldots, (n_i + 1), \ldots\rangle. \tag{6.5}$$

Since each operator affects only one of the occupation numbers, it is easy to show that operators for different k_i commute, while those for the same k_i satisfy (5.59). In summary, the commutation relations are

$$[\hat{a}(k_i), \hat{a}(k_j)] = [\hat{a}^\dagger(k_i), \hat{a}^\dagger(k_j)] = 0 \tag{6.6}$$

$$[\hat{a}(k_i), \hat{a}^\dagger(k_j)] = \delta_{ij}. \tag{6.7}$$

If some of the k variables are continuous, we must revert to the previous representation. The commutation relation (6.7) becomes

$$[\hat{a}(k), \hat{a}^\dagger(k')] = \delta(k - k'). \tag{6.8}$$

If we restrict attention to basis vectors whose k arguments are all different, then the action of the creation and annihilation operators is

$$\hat{a}^\dagger(k)|k_1, k_2, \ldots, k_N\rangle = |k_1, k_2, \ldots, k_N, k\rangle \tag{6.9}$$

$$\hat{a}(k)|k_1, k_2, \ldots, k_N\rangle = \sum_{n=1}^{N} \delta(k - k_n)|k_1, k_2, \ldots, (k_n), \ldots, k_N\rangle \tag{6.10}$$

where, in the second equation, (k_n) denotes a label that is missing from the original list. By acting with $[\hat{a}(k), \hat{a}^\dagger(k')]$ on an arbitrary basis vector, it is easily verified that (6.9) and (6.10) imply the relation (6.8). Readers will also find it instructive to verify from the above equations that $\hat{a}^\dagger(k)$ is indeed the adjoint of $\hat{a}(k)$.

The entire set of basis vectors can be constructed by the method we used in the case of the harmonic oscillator. We start from the vacuum state $|0\rangle$, which contains no particles, and use the creation operator to populate it:

$$|k_1, k_2, \ldots, k_N\rangle = \hat{a}^\dagger(k_N) \cdots \hat{a}^\dagger(k_2)\hat{a}^\dagger(k_1)|0\rangle. \tag{6.11}$$

The Hilbert space constructed in this way is called a *Fock space*. A subtle point is worth noting. When particles interact with each other, it is still possible to form state vectors in terms of single-particle states, but these will not, in general, be energy eigenstates. It is not necessarily true that every possible state of the system can be represented as a superposition of the Fock basis vectors, so the Fock space constructed according to (6.11) may be only a part of the whole Hilbert space. For many purposes, though, it will not be necessary to worry about this.

6.2 Field Operators and Observables

From now on, we always take the single-particle states to be momentum eigenstates. For the moment, we consider only spinless particles, so k stands just for the three momentum components, or rather for the wavevector $\boldsymbol{k} = \boldsymbol{p}/\hbar$. The wavefunction for a single particle in the state $|\Psi(t)\rangle$ can be written as

$$\begin{aligned}\Psi(\boldsymbol{x}, t) = \langle \boldsymbol{x}|\Psi(t)\rangle &= (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\langle \boldsymbol{k}|\Psi(t)\rangle \\ &= (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\langle 0|\hat{a}(k)|\Psi(t)\rangle. \qquad (6.12)\end{aligned}$$

The annihilation operator $\hat{a}(k)$ creates the one-particle bra vector from the vacuum because it is the adjoint of $\hat{a}^\dagger(k)$. In the non-relativistic theory, we define the Schrödinger-picture *field operators* by

$$\hat{\psi}(\boldsymbol{x}) = (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\hat{a}(k) \qquad (6.13)$$

$$\hat{\psi}^\dagger(\boldsymbol{x}) = (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\hat{a}^\dagger(k). \qquad (6.14)$$

Obviously, these create or annihilate a particle at a definite point $\boldsymbol{x}$, rather than in a state of definite momentum; for example, $|\boldsymbol{x}\rangle = \hat{\psi}^\dagger(\boldsymbol{x})|0\rangle$. In relativistic theories, we shall find that the situation is a little more complicated because of the need to maintain Lorentz covariance. The commutation relations for the field operators follow from those of $\hat{a}(k)$ and $\hat{a}^\dagger(k)$. They are

$$[\hat{\psi}(\boldsymbol{x}), \hat{\psi}(\boldsymbol{x}')] = [\hat{\psi}^\dagger(\boldsymbol{x}), \hat{\psi}^\dagger(\boldsymbol{x}')] = 0 \qquad (6.15)$$

$$[\hat{\psi}(\boldsymbol{x}), \hat{\psi}^\dagger(\boldsymbol{x}')] = \delta(\boldsymbol{x} - \boldsymbol{x}'). \qquad (6.16)$$

The operators that represent observable properties of many-particle systems are constructed from the creation and annihilation operators or from the field operators. The operator $\hat{n}(k) = \hat{a}^\dagger(k)\hat{a}(k)$ is a number operator, which counts the number of particles in the state k, if k is discrete. If the momentum takes a continuous range of values, then $\hat{n}(k)\mathrm{d}^3k$ counts the number of particles in the momentum range d^3k near k. The total number of particles is counted by the operator

$$\hat{N} = \int \mathrm{d}^3k\, \hat{a}^\dagger(k)\hat{a}(k) = \int \mathrm{d}^3x\, \hat{\psi}^\dagger(\boldsymbol{x})\hat{\psi}(\boldsymbol{x}) \qquad (6.17)$$

and, by summing $\hbar\boldsymbol{k}$ times the number of particles having that momentum, we find that the total momentum is represented by the operator

$$\hat{\boldsymbol{P}} = \int \mathrm{d}^3k\,(\hbar\boldsymbol{k})\hat{a}^\dagger(k)\hat{a}(k) = \int \mathrm{d}^3x\,\hat{\psi}^\dagger(\boldsymbol{x})(-\mathrm{i}\hbar\boldsymbol{\nabla})\hat{\psi}(\boldsymbol{x}). \tag{6.18}$$

The number and total momentum are *one-body* operators, in the sense that they represent the total for the system of a property possessed by individual particles. Kinetic energy, mass, electric charge and the potential energy due to an externally applied field are examples of other properties of the same kind. There is clearly a general rule for constructing one-body operators. If $A(\boldsymbol{x}, -\mathrm{i}\hbar\boldsymbol{\nabla})$ is the wave-mechanical operator that represents some property of a single particle, then the total property for the whole system is represented by

$$\hat{A} = \int \mathrm{d}^3x\,\hat{\psi}^\dagger(\boldsymbol{x})A(\boldsymbol{x}, -\mathrm{i}\hbar\boldsymbol{\nabla})\hat{\psi}(\boldsymbol{x}). \tag{6.19}$$

We may also consider operators that depend for their definition on two or more particles at a time. An example is the Coulomb potential, which acts between two particles. In a state with particles at the points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, the total potential energy is

$$\begin{aligned} V &= \tfrac{1}{2}\sum_{i,j=1}^{N} V(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \tfrac{1}{2}\int \mathrm{d}^3x\,\mathrm{d}^3x'\,V(\boldsymbol{x}, \boldsymbol{x}')\sum_{i,j=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i)\delta(\boldsymbol{x}'-\boldsymbol{x}_j) \end{aligned} \tag{6.20}$$

the terms with $i = j$ being excluded from the sum. This will be correctly represented if we can find an operator which, when acting on any state of the form $|\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\rangle$, gives the same state multiplied by the sum of delta functions in (6.20). The action of the field operators on this state is exactly analogous to (6.9) and (6.10), and I leave it as an exercise for readers to verify that the total potential energy is represented by the operator

$$\hat{V} = \tfrac{1}{2}\int \mathrm{d}^3x\,\mathrm{d}^3x'\,\hat{\psi}^\dagger(\boldsymbol{x})\hat{\psi}^\dagger(\boldsymbol{x}')V(\boldsymbol{x}, \boldsymbol{x}')\hat{\psi}(\boldsymbol{x}')\hat{\psi}(\boldsymbol{x}). \tag{6.21}$$

6.3 Equation of Motion and Lagrangian Formalism for Field Operators

We have dealt so far only with Schrödinger-picture field operators. In the Heisenberg picture, time-dependent operators are defined by the usual method through (5.35):

$$\hat{\psi}(\boldsymbol{x}, t) = \mathrm{e}^{\mathrm{i}\hat{H}t/\hbar}\hat{\psi}(\boldsymbol{x})\mathrm{e}^{-\mathrm{i}\hat{H}t/\hbar}. \tag{6.22}$$

For a system of free particles, the Hamiltonian is just the kinetic energy operator. Because the Hamiltonian commutes with itself, it can be expressed in terms of either the Schrödinger-picture or the Heisenberg-picture fields:

$$\hat{H} = \int \mathrm{d}^3x\, \hat{\psi}^\dagger(\boldsymbol{x}) \left(-\frac{\hbar^2}{2m}\nabla^2 \right) \hat{\psi}(\boldsymbol{x}) = \int \mathrm{d}^3x\, \hat{\psi}^\dagger(\boldsymbol{x},t) \left(-\frac{\hbar^2}{2m}\nabla^2 \right) \hat{\psi}(\boldsymbol{x},t). \tag{6.23}$$

Readers to whom this is not obvious should verify it by substituting (6.22) into the second expression. The same is true if $\hat{H}$ contains a potential energy of the form (6.21) or a one-body external potential. The Heisenberg-picture operators satisfy the commutation relations (6.15) and (6.16), provided that all operators are evaluated at the same time; these are called *equal-time commutation relations*. By using them in (5.36), we can find the equation of motion for $\hat{\psi}(\boldsymbol{x},t)$. If we include the potential (6.21) and an external potential $U(\boldsymbol{x})$, the result is

$$\begin{aligned} \mathrm{i}\hbar\frac{\partial}{\partial t}\hat{\psi}(\boldsymbol{x},t) &= -\frac{\hbar^2}{2m}\nabla^2\hat{\psi}(\boldsymbol{x},t) + U(\boldsymbol{x})\hat{\psi}(\boldsymbol{x},t) \\ &\quad + \int \mathrm{d}^3x'\, \hat{\psi}^\dagger(\boldsymbol{x}',t)V(\boldsymbol{x},\boldsymbol{x}')\hat{\psi}(\boldsymbol{x}',t)\hat{\psi}(\boldsymbol{x},t). \end{aligned} \tag{6.24}$$

When the two-body potential is absent, this is the same as the Schrödinger equation satisfied by the wavefunction. This is just as well, since the single-particle wavefunction (6.12) can be written in the Heisenberg picture as $\Psi(\boldsymbol{x},t) = \langle 0|\hat{\psi}(\boldsymbol{x},t)|\Psi\rangle$, and it must obey the Schrödinger equation.

I shall now show that the whole structure of second quantization can be obtained from a Lagrangian formalism, by means of the canonical quantization prescription described in chapter 5. For brevity, I shall give the derivation just for the free-particle theory whose Hamiltonian is (6.23), but readers should be able to extend it without difficulty to the case of particles interacting through a two-body potential. Consider the action defined by

$$S = \int \mathrm{d}t\, \mathrm{d}^3x\, \psi^*(\boldsymbol{x},t) \left(\mathrm{i}\hbar\frac{\partial}{\partial t} + \frac{\hbar^2}{2m}\nabla^2 \right) \psi(\boldsymbol{x},t) \tag{6.25}$$

where $\psi(\boldsymbol{x},t)$ is a complex function, not, for the moment, a field operator. In chapter 3, we saw that Maxwell's equations for the electromagnetic field could be obtained by finding the Euler–Lagrange equations for an action somewhat akin to this. In this case, the real and imaginary parts of ψ are independent functions, but it is more convenient to treat ψ and ψ^* as the independent variables. By varying $\psi^*(\boldsymbol{x},t)$ we obtain Schrödinger's equation for $\psi(\boldsymbol{x},t)$ and by varying $\psi(\boldsymbol{x},t)$ itself we get the complex conjugate of the same equation. The values of $\psi(\boldsymbol{x},t)$ at each point $\boldsymbol{x}$ form an infinite set of generalized coordinates, and there is an infinite set of conjugate momenta, which form a function $\Pi(\boldsymbol{x},t)$. This function is found by functional differentiation (which is explained in appendix A for readers who are not familiar with it):

$$\Pi(\boldsymbol{x},t) = \frac{\delta S}{\delta\dot{\psi}(\boldsymbol{x},t)} = \mathrm{i}\hbar\psi^*(\boldsymbol{x},t). \tag{6.26}$$

In (6.25), the time derivative acts only on ψ or, if we integrate by parts, only on ψ^*. Therefore, we cannot define conjugate momenta for ψ and ψ^* at the same time. This kind of difficulty was mentioned in chapter 5, and the solution that works here is to ignore ψ^* as an independent variable, except for the purpose of deriving the equation of motion. Then the Hamiltonian is

$$H = \int \mathrm{d}^3x\, \Pi\dot{\psi} - L = \int \mathrm{d}^3x\, \psi^* \left(-\frac{\hbar^2}{2m}\nabla^2\right)\psi. \tag{6.27}$$

To get back to our quantum theory, we simply follow the canonical quantization scheme, replacing $\psi(\boldsymbol{x},t)$ with the field operator $\hat{\psi}(\boldsymbol{x},t)$ and its complex conjugate with $\hat{\psi}^\dagger(\boldsymbol{x},t)$. The Hamiltonian (6.27) becomes identical to (6.23). In the canonical commutator (5.49), the coordinate $\hat{q}_i$ becomes $\hat{\psi}(\boldsymbol{x},t)$, the momentum $\hat{p}_j$ is replaced with the momentum $\hat{\Pi}(\boldsymbol{x}',t)$ obtained from (6.26), and the Kronecker symbol is replaced by $\delta(\boldsymbol{x}-\boldsymbol{x}')$. The result is none other than the commutator (6.16) for the field operators. For the kind of theory we have been considering, this new bit of formalism provides no new information, since we have just returned to our starting point. Suppose, however, that we wish to treat the electromagnetic field as a quantum system. The analysis we have just been through shows us how to do this, although there is an added difficulty to be overcome, as will be discussed in chapter 9. The vector potential A^μ becomes a field operator, which obeys Maxwell's equations rather than the Schrödinger equation, and its commutation relations will again be given by the canonical prescription. In the light of our experience in this chapter, we may anticipate that this field operator can be interpreted in terms of creation and annihilation operators for particles, namely *photons*, which are quanta of electromagnetic energy. In fact, the Lagrangian formalism provides the most convenient basis for most relativistic field theories.

6.4 Second Quantization for Fermions

Many of the important applications of non-relativistic field theory concern electronic systems. Electrons have spin $\frac{1}{2}$ and are therefore fermions. Although the consequences of this are far reaching, the modifications needed in the basic theory are quite simple. We must take the label k of single-particle states to include the variable s, which measures the component of spin along a chosen quantization axis and has the values $\pm\frac{1}{2}$. Slightly more tricky is the antisymmetry of multiparticle states. For simplicity, let us consider two-particle states, for which $|k,k'\rangle = -|k',k\rangle$. I shall follow the common practice of using $\hat{b}(k)$ and $\hat{b}^\dagger(k)$ to denote fermionic annihilation and creation operators, to distinguish them from bosonic ones. It is now important to keep track of the ordering of k labels in a state vector. A sensible convention when using $\hat{b}^\dagger(k)$ to add a particle is to place the label for the added particle at the end of the list. Thus,

$$\hat{b}^\dagger(k')\hat{b}^\dagger(k)|0\rangle = |k, k'\rangle = -\hat{b}^\dagger(k)\hat{b}^\dagger(k')|0\rangle. \tag{6.28}$$

Similarly, the annihilation operator can be regarded as removing the last particle in the list. It can, of course, remove any particle in the state, so to write down the result we first move the particle to the end, if necessary, incurring a minus sign for each interchange of particle labels. For a two-particle state,

$$\hat{b}(k)|k_1, k_2\rangle = \delta(k - k_2)|k_1\rangle - \delta(k - k_1)|k_2\rangle = -\hat{b}(k)|k_2, k_1\rangle. \tag{6.29}$$

More generally, (6.10) is modified to read

$$\hat{b}(k)|k_1, k_2, \ldots, k_N\rangle = \sum_{n=1}^{N}(-1)^{N-n}\delta(k-k_n)|k_1, k_2, \ldots, (k_n), \ldots, k_N\rangle. \tag{6.30}$$

Evidently, $\hat{b}(k)$ and $\hat{b}^\dagger(k)$ cannot obey the commutation relations (6.6)-(6.8). In fact, as it is not difficult to see, the relations consistent with the antisymmetry of the state vectors are the *anticommutation relations*

$$\{\hat{b}(k), \hat{b}(k')\} = \{\hat{b}^\dagger(k), \hat{b}^\dagger(k')\} = 0 \tag{6.31}$$
$$\{\hat{b}(k), \hat{b}^\dagger(k')\} = \delta(k - k') \tag{6.32}$$

where the anticommutator is defined by $\{\hat{A}, \hat{B}\} = \hat{A}\hat{B} + \hat{B}\hat{A}$. In particular, this means that $\hat{b}^\dagger(k)\hat{b}^\dagger(k) = 0$. Acting twice with the same creation operator gives zero, instead of two particles in the same state. This, of course, is the second-quantization version of the Pauli exclusion principle, which asserts that no two identical fermions can occupy the same single-particle state.

Field operators for fermions can be constructed in the same way as for bosons, except that we have to take account of spin polarization. Since k now stands for $(\boldsymbol{k}, s)$, the definitions (6.13) and (6.14) become

$$\hat{\psi}_s(\boldsymbol{x}) = (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\hat{b}(\boldsymbol{k}, s) \tag{6.33}$$

$$\hat{\psi}_s^\dagger(\boldsymbol{x}) = (2\pi)^{-3/2}\int \mathrm{d}^3k\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\hat{b}^\dagger(\boldsymbol{k}, s). \tag{6.34}$$

For example, $\hat{\psi}_s^\dagger(\boldsymbol{x})$ creates a particle at the point $\boldsymbol{x}$ with spin polarization s. The anticommutation relations that replace (6.15) and (6.16) are

$$\{\hat{\psi}_s(\boldsymbol{x}), \hat{\psi}_{s'}(\boldsymbol{x}')\} = \{\hat{\psi}_s^\dagger(\boldsymbol{x}), \hat{\psi}_{s'}^\dagger(\boldsymbol{x}')\} = 0 \tag{6.35}$$
$$\{\hat{\psi}_s(\boldsymbol{x}), \hat{\psi}_{s'}^\dagger(\boldsymbol{x}')\} = \delta_{ss'}\delta(\boldsymbol{x} - \boldsymbol{x}'). \tag{6.36}$$

These are all the changes we need to make in order to accommodate fermions. In equations (6.19), (6.21) and (6.24), it is necessary only to add spin labels to the fields and include a sum over these labels with each space integration. I have ordered the operators in these expressions so as to make them correct for both fermions and bosons.

Exercises

6.1. Let $A(\boldsymbol{x}, -\mathrm{i}\hbar\boldsymbol{\nabla})$, $B(\boldsymbol{x}, -\mathrm{i}\hbar\boldsymbol{\nabla})$ and $C(\boldsymbol{x}, -\mathrm{i}\hbar\boldsymbol{\nabla})$, be wave-mechanical operators with the commutation relation $[A, B] = C$. Show that the corresponding second-quantized one-body operators $\hat{A}$, $\hat{B}$ and $\hat{C}$ satisfy the same commutation relation, if the field operators have either the commutation relations (6.15) and (6.16) appropriate to bosons or the anticommutation relations (6.35) and (6.36) appropriate to fermions.

6.2. Using time-independent field operators, show that the Hamiltonian (6.23) can be expressed as

$$\hat{H} = \int \mathrm{d}^3x \; \hbar\omega(\boldsymbol{k})\hat{a}^\dagger(k)\hat{a}(k)$$

where $\omega(\boldsymbol{k}) = \hbar k^2/2m$. Show that for any n, $\hat{H}^n\hat{a}(\boldsymbol{k}) = \hat{a}(\boldsymbol{k})[\hat{H} - \hbar\omega(\boldsymbol{k})]^n$ and hence that the time-dependent field operator (6.22) is

$$\hat{\psi}(\boldsymbol{x}, t) = (2\pi)^{-3/2}\int \mathrm{d}^3k \; \exp[\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x} - \mathrm{i}\omega(\boldsymbol{k})t]\hat{a}(k).$$

Check that this works for both bosons and fermions. There is no such simple expression for $\hat{\psi}(\boldsymbol{x}, t)$ if the particles interact.

Chapter 7

Relativistic Wave Equations and Field Theories

Up to this point, our study of quantum mechanics has concerned itself with the behaviour of particles that inhabit a Galilean spacetime. For many purposes, in atomic, molecular and condensed matter physics, this theory is quite adequate. We saw in earlier chapters, however, that our actual spacetime has a structure which is much closer to that of the Minkowski spacetime of special relativity and that more general structures must be considered when gravitational phenomena are significant. From a purely theoretical point of view, it is therefore important to formulate quantum theory in a way which is consistent with these more general spacetimes. The benefits of constructing a relativistic quantum theory actually go far beyond the aesthetic satisfaction of making our geometrical and quantum-mechanical reasoning compatible. For one thing, we shall discover that the relativistic theory provides a deeper understanding of spin and the distinction between fermions and bosons, which in the non-relativistic theory appear simply as facts of life that we must strive to accommodate. Also, of course, there are many situations in which relativistic effects become observable, for which non-relativistic theory provides no explanation. The most obvious are high-energy scattering experiments, in which particles acquire kinetic energies comparable with or greater than their rest energies mc^2, and the correct 4-momentum (3.34) must be used. There are, however, more subtle effects, such as the spin-orbit coupling that is essential for interpreting atomic spectra, which are also of relativistic origin.

For the most part, I shall deal only with quantum theory in Minkowski spacetime, which is well understood. At the end of the chapter, I shall deal rather more briefly with the question of setting up quantum theories in curved spacetimes, which involves some surprising difficulties and is not quite so well understood. If our world is thoroughly quantum-mechanical (and the prevailing view is that it must be), then we ought to treat the geometrical structures themselves in quantum-mechanical terms, which means constructing a quantum theory of gravity. Attempts to deal with the metric tensor (or, perhaps, the affine connection) by the methods to be dealt with in this chapter have generally not been successful, though it is not entirely clear that no such theory is possible. The majority view among physicists who study such matters is that string theory offers the best hope of a theory of gravity that is consistent with quantum mechanics, and I shall have something to say about this in chapter 15.

From now on, I shall write all equations having to do with relativistic theories in terms of *natural units*, which are defined so that $\hbar = c = 1$. This leaves us free to define one fundamental unit, which is normally taken to be energy, measured, say, in MeV. In these units, length and time have the same dimensions and are measured in $(\mathrm{MeV})^{-1}$. Mass, momentum and energy have the same dimensions, being measured in MeV. Appendix C discusses these units in more detail and gives some conversion factors between natural and laboratory units.

7.1 The Klein–Gordon Equation

If we wish to invent a Minkowski-spacetime version of wave mechanics, the first problem to be overcome is that the Schrödinger equation (5.7) expresses the non-relativistic relationship between energy and momentum. The relationship in special relativity is that implied by (3.34), which may be written in various ways as

$$E^2 - \boldsymbol{p}\cdot\boldsymbol{p} = (p^0)^2 - \boldsymbol{p}\cdot\boldsymbol{p} = p_\mu p^\mu = m^2. \tag{7.1}$$

At least for free particles, it is a simple matter to convert this into a relativistic wave equation, called the *Klein–Gordon equation*. We just substitute the differential operators (5.5) and (5.6) and let the resulting operator act on a wavefunction $\phi(\boldsymbol{x}, t)$:

$$\left[\frac{\partial^2}{\partial t^2} - \nabla^2 + m^2\right]\phi(\boldsymbol{x}, t) = \left[\partial_\mu\partial^\mu + m^2\right]\phi(x) = \left[\Box + m^2\right]\phi(x) = 0. \tag{7.2}$$

The d'Alembertian operator $\Box$ defined here is the Minkowskian version of the Laplacian ∇^2; it is sometimes written as $\Box^2$. We should certainly expect the Klein–Gordon equation to be valid for a free relativistic particle, but whether it should be regarded as a generalization of the Schrödinger equation is a moot point, since it is not related in a simple way to a time-evolution equation of the form (5.32).

It is important to ask how the wavefunction is to be interpreted in a relativistic context. If (7.2) is to have a Lorentz covariant meaning, then ϕ must be some kind of 4-tensor, as discussed in §3.5. For now, we shall consider spinless particles whose wavefunctions have only a single component, so ϕ must be a scalar. This implies that the probability density is not correctly given by (5.4). Like the number density in (3.36), it must be the time-like component of a conserved 4-vector, whose other components are the probability current density. In a loose, intuitive manner, we can think of the probability density as a kind of number density. In the non-relativistic theory, the current density is

$$\boldsymbol{j}(\boldsymbol{x}, t) = \frac{1}{m}\mathrm{Re}[\Psi^*(-\mathrm{i}\boldsymbol{\nabla})\Psi] = -\frac{\mathrm{i}}{2m}\Psi^*\overleftrightarrow{\boldsymbol{\nabla}}\Psi \tag{7.3}$$

where the notation $A\overleftrightarrow{\boldsymbol{\nabla}}B$ means $A\boldsymbol{\nabla}B - (\boldsymbol{\nabla}A)B$. Intuitively, this expression is rather like (velocity × density), as in (3.37). More precisely, the reason for (7.3) is that it satisfies the equation of continuity $\partial P/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{j} = 0$, as may easily be verified by using the Schrödinger equation and its complex conjugate. In the present case, it may be similarly verified using the Klein–Gordon equation that the equation of continuity in the form $\partial_\mu j^\mu = 0$ is satisfied, provided that we identify the 4-vector probability current density as

$$j^\mu(x) = \frac{\mathrm{i}}{2m}\phi^*\overleftrightarrow{\partial}^\mu\phi. \tag{7.4}$$

This is fortunate insofar as (7.4) is manifestly a 4-vector, so that the equation of continuity is Lorentz covariant. The unfortunate thing about (7.4) is that j^0, which we want to identify as the probability density, is, unlike $|\Psi|^2$, not necessarily positive. This is one of two problems that afflict all relativistic wave equations.

The second problem emerges when we look at plane-wave solutions of the Klein–Gordon equation. Evidently, the function

$$\phi_k(x) = \exp(-\mathrm{i}k\cdot x) \tag{7.5}$$

where $k\cdot x = k^0t - \boldsymbol{k}\cdot\boldsymbol{x}$, is a solution of (7.2) and also an energy–momentum eigenfunction, provided that $k^0 = \pm(\boldsymbol{k}^2 + m^2)^{1/2}$. The negative-energy solutions are a severe embarrassment, because they imply the existence of single-particle states with energy less than that of the vacuum. Intuitively, this is nonsensical. In fact, there is no lower limit to the energy spectrum. This means that the vacuum is unstable, since an infinite amount of energy could be released from it by the spontaneous creation of particles in negative-energy states. We can see that this problem is related to the first one, because it is the negative-energy states that give rise to a negative probability density.

Because of these problems, the Klein–Gordon equation does not lead to a tenable wave-mechanical theory of relativistic particles. It is, indeed, impossible to construct such a theory. We shall see shortly that it does lead to a perfectly sensible quantum field theory. To develop this field theory, we follow the canonical quantization procedure explained in §6.3, but the requirement of Lorentz covariance leads to some minor changes. Like the Schrödinger equation, the Klein–Gordon equation can be obtained as an Euler–Lagrange equation from an action. Assuming that ϕ is a complex function, the action is

$$S = \int \mathrm{d}^4x\,\mathcal{L}(\phi) \tag{7.6}$$

where the Lagrangian density is given by

$$\mathcal{L}(\phi) = (\partial_\mu \phi^*)(\partial^\mu \phi) - m^2 \phi^* \phi. \tag{7.7}$$

This action is manifestly a scalar quantity, as we require for a Lorentz covariant theory. In contrast to the non-relativistic theory, it contains the time derivatives of both ϕ and ϕ^*, so two independent canonical momenta can be defined:

$$\Pi(x) = \partial^0 \phi^*(x) \qquad \Pi^*(x) = \partial^0 \phi(x). \tag{7.8}$$

The general solution of the Klein–Gordon equation can be written in terms of energy–momentum eigenfunctions. To ensure that it is a scalar, we first write it in a form that does not distinguish between space and time components of the energy–momentum 4-vector:

$$\phi(x) = \int \frac{\mathrm{d}^4 k}{(2\pi)^3}\, \delta(k^2 - m^2)\alpha(k)\mathrm{e}^{-\mathrm{i}k\cdot x}. \tag{7.9}$$

The energy (k^0) integral can be carried out using the delta function. We get two terms, corresponding to the positive and negative energy solutions, with $k^0 = \pm\omega(\boldsymbol{k})$, where $\omega(\boldsymbol{k}) = (\boldsymbol{k}^2 + m^2)^{1/2}$. For reasons that will become apparent, we write the coefficient $\alpha(k)$ as

$$\alpha(k) = \begin{cases} a(\boldsymbol{k}) & \text{for } k^0 = +\omega(\boldsymbol{k}) \\ c^*(-\boldsymbol{k}) & \text{for } k^0 = -\omega(\boldsymbol{k}). \end{cases} \tag{7.10}$$

Then, after changing the sign of $\boldsymbol{k}$ in the negative-energy term, we get

$$\phi(x) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3 2\omega(\boldsymbol{k})} \left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}k\cdot x} + c^*(\boldsymbol{k})\mathrm{e}^{\mathrm{i}k\cdot x} \right]. \tag{7.11}$$

In each term, k^0 now stands for $+\omega(\boldsymbol{k})$. The $2\omega(\boldsymbol{k})$ in the denominator appears for the reason explained in appendix A. Because of this factor, the coefficients $a(\boldsymbol{k})$ and $c(\boldsymbol{k})$ cannot be obtained by a simple Fourier transformation. Instead, we have the expressions

$$a(\boldsymbol{k}) = \mathrm{i} \int \mathrm{d}^3 x\, \mathrm{e}^{\mathrm{i}k\cdot x} \overleftrightarrow{\partial}^0 \phi(x) = \int \mathrm{d}^3 x\, \mathrm{e}^{\mathrm{i}k\cdot x} \left[\omega(\boldsymbol{k})\phi(x) + \mathrm{i}\Pi^*(x) \right] \tag{7.12}$$

$$c(\boldsymbol{k}) = \mathrm{i} \int \mathrm{d}^3 x\, \mathrm{e}^{\mathrm{i}k\cdot x} \overleftrightarrow{\partial}^0 \phi^*(x) = \int \mathrm{d}^3 x\, \mathrm{e}^{\mathrm{i}k\cdot x} \left[\omega(\boldsymbol{k})\phi^*(x) + \mathrm{i}\Pi(x) \right] \tag{7.13}$$

which have a rather similar structure to the energy-lowering operator (5.56) for the harmonic oscillator. With these expressions in hand, we are ready to develop the second-quantized description.

7.2 Scalar Field Theory for Free Particles

As in the non-relativistic case, we carry out second quantization by replacing complex functions with field operators. Because a relativistic theory treats space and time on much the same footing, these are initially given as time-dependent Heisenberg-picture operators. Nevertheless, we are still free only to specify the equal-time commutators. First of all, we have

$$\left[\hat{\phi}(\boldsymbol{x},t),\hat{\Pi}(\boldsymbol{x}',t)\right]=\mathrm{i}\delta(\boldsymbol{x}-\boldsymbol{x}') \tag{7.14}$$

$$\left[\hat{\phi}(\boldsymbol{x},t),\hat{\phi}(\boldsymbol{x}',t)\right]=\left[\hat{\Pi}(\boldsymbol{x},t),\hat{\Pi}(\boldsymbol{x}',t)\right]=0. \tag{7.15}$$

Taking the adjoints of these equations, we find that $\hat{\phi}^\dagger$ and $\hat{\Pi}^\dagger$ satisfy exactly the same relations. The two sets of operators $\{\hat{\phi},\hat{\Pi}\}$ and $\{\hat{\phi}^\dagger,\hat{\Pi}^\dagger\}$ are to be treated as independent variables, so we also have

$$\left[\hat{\phi}(\boldsymbol{x},t),\hat{\phi}^\dagger(\boldsymbol{x}',t)\right]=\left[\hat{\Pi}(\boldsymbol{x},t),\hat{\Pi}^\dagger(\boldsymbol{x}',t)\right]=\left[\hat{\phi}(\boldsymbol{x},t),\hat{\Pi}^\dagger(\boldsymbol{x}',t)\right]=0. \tag{7.16}$$

By using these commutators, we can work out the commutation relations for the operator versions of $a(\boldsymbol{k})$ and $c(\boldsymbol{k})$ from (7.12) and (7.13). The result is that

$$\left[\hat{a}(\boldsymbol{k}),\hat{a}^\dagger(\boldsymbol{k}')\right]=\left[\hat{c}(\boldsymbol{k}),\hat{c}^\dagger(\boldsymbol{k}')\right]=(2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k}-\boldsymbol{k}') \tag{7.17}$$

while all other commutators between these operators are zero. Apart from the normalization factor $(2\pi)^3 2\omega(\boldsymbol{k})$, we recognize these as two independent sets of creation and annihilation operators, similar to those in (6.8). The effect of the normalization factor for single-particle states is that

$$\langle\boldsymbol{k}|\boldsymbol{k}'\rangle=(2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k}-\boldsymbol{k}'). \tag{7.18}$$

This is a Lorentz-covariant normalization, as we can see by constructing the corresponding wavefunction. To do this, we need the vector $|\boldsymbol{x}\rangle$, which must be given by an expression similar to (5.70). The exact expression is

$$|\boldsymbol{x}\rangle=\int\frac{\mathrm{d}^3k}{(2\pi)^3\sqrt{2\omega(\boldsymbol{k})}}\,\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}|\boldsymbol{k}\rangle \tag{7.19}$$

in which the factor of $\sqrt{2\omega(\boldsymbol{k})}$ is required to get the correct orthonormality relation $\langle\boldsymbol{x}|\boldsymbol{x}'\rangle=\delta(\boldsymbol{x}-\boldsymbol{x}')$. Then the wavefunction

$$\psi_k(\boldsymbol{x})=\langle\boldsymbol{x}|\boldsymbol{k}\rangle=\sqrt{2\omega(\boldsymbol{k})}\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \tag{7.20}$$

gives a probability density $P(\boldsymbol{x})=|\psi_k(\boldsymbol{x})|^2=2\omega(\boldsymbol{k})$. Loosely, this corresponds to $2\omega(\boldsymbol{k})$ particles per unit volume. Under a Lorentz transformation, it transforms as the time-like component of a 4-vector, as it ought to.

The fact that we have two sets of creation and annihilation operators leads to the resolution of the problems of negative energies and probabilities. The field theory we have constructed actually describes two species of particles; particles of one species are called the *antiparticles* of the other. For the sake of argument, I shall refer to the particles created by $\hat{a}^\dagger$ as 'particles' and to those created by $\hat{c}^\dagger$ as 'antiparticles', though the theory itself does not care which is which. The solution to the problem of negative energies is apparent from (7.11) when we reinterpret it as a field operator. The coefficient of the positive-energy wavefunction $\mathrm{e}^{-\mathrm{i}k\cdot x}$ is, as in the non-relativistic theory, the annihilation operator for particles. However, the coefficient of the negative-energy wavefunction $\mathrm{e}^{\mathrm{i}k\cdot x}$ is not an annihilation operator for particles in negative-energy states, but rather a creation operator for positive-energy antiparticles. We can construct the Hamiltonian operator by the usual canonical method. It is

$$\begin{aligned}\hat{H} &= \int \mathrm{d}^3x \left[\frac{\partial\hat{\phi}^\dagger}{\partial t}\frac{\partial\hat{\phi}}{\partial t} + (\nabla\hat{\phi}^\dagger)\cdot(\nabla\hat{\phi}) + m^2\hat{\phi}^\dagger\hat{\phi}\right]\\ &= \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})}\,\omega(\boldsymbol{k})\left[\hat{a}^\dagger(\boldsymbol{k})\hat{a}(\boldsymbol{k}) + \hat{c}^\dagger(\boldsymbol{k})\hat{c}(\boldsymbol{k}) + (2\pi)^3 2\omega(\boldsymbol{k})\delta(0)\right].\end{aligned} \tag{7.21}$$

In the second expression, the last term comes from rewriting $\hat{c}(\boldsymbol{k})\hat{c}^\dagger(\boldsymbol{k})$ by means of the commutator. If we act on the vacuum state, which contains no particles or antiparticles, the first two terms give zero. The last term is an infinite constant. It may be dropped on the usual grounds that the total energy of a system is defined only up to an arbitrary constant, and the most sensible choice for the energy of the vacuum is zero. (If we allow the structure of spacetime to be determined by Einstein's equations (4.17), however, the energy of the vacuum contributes to $T^{\mu\nu}$ and must be considered more carefully.) Another way of looking at this is to remember that the ordering of operators is not unambiguously prescribed by the quantization procedure. We can regard the vanishing of the vacuum energy as a criterion for ordering operators such that annihilation operators appear to the right of creation operators. This is called *normal ordering*. Bearing in mind the normalization in (7.17), we recognize (7.21) as summing the quantity

(energy of state $\boldsymbol{k}$) × (number of particles and antiparticles in state $\boldsymbol{k}$)

over positive-energy states. Thus, the total energy is positive.

The solution of the problem of negative probabilities is quite similar. We define a number operator $\hat{N}$ by integrating over all space the operator corresponding to the probability density in (7.4):

$$\begin{aligned}\hat{N} = \int \mathrm{d}^3x :\hat{j}^0: &= \mathrm{i}\int \mathrm{d}^3x :\hat{\phi}^\dagger(\boldsymbol{x},t)\overleftrightarrow{\partial}^0\hat{\phi}(\boldsymbol{x},t):\\ &= \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})}\left[\hat{a}^\dagger(\boldsymbol{k})\hat{a}(\boldsymbol{k}) - \hat{c}^\dagger(\boldsymbol{k})\hat{c}(\boldsymbol{k})\right]\end{aligned} \tag{7.22}$$

where the colons $:\cdots:$ denote normal ordering of the creation and annihilation operators. Again, the factor $1/2\omega(\boldsymbol{k})$ appears just because of the covariant normalization and in effect replaces the $1/2m$ in (7.4). We see that $\hat{N}$ represents the quantity

$$\text{(number of particles)} - \text{(number of antiparticles)}.$$

A negative value for this quantity simply indicates a state with more antiparticles than particles and presents no difficulty. Another way of expressing this is to assign to each particle a *particle number* $n = 1$ and to each antiparticle a particle number $n = -1$. Then $\hat{N}$ may be said to represent the net particle number, rather than the number of particles. In the field operator obtained from (7.11), both terms act on a given state to reduce the particle number by one unit, either by annihilating a particle or by creating an antiparticle. This rule applies to any other properties that the particles may possess (except that the masses of particles and antiparticles are identical). For example, if the particles carry an electric charge, then their antiparticles carry exactly the opposite charge, and the same is true of all the other *quantum numbers* (lepton number, baryon number, isospin, strangeness, etc) which are required to classify the observed particles. Historically, the existence of antiparticles was predicted by Dirac (1928, 1929) on the basis of his relativistic wave equation for electrons discussed in the next section, and the antielectron, or positron, was discovered experimentally by Anderson (1933) in cosmic ray showers. All observed particles are indeed found to have antiparticles. However, particles and antiparticles may in some cases be identical. Mathematically, this will be so if $\hat{c}(\boldsymbol{k}) = \hat{a}(\boldsymbol{k})$, which means that the wavefunction (7.11) is real and the corresponding field operator is Hermitian. In that case, the number operator is identically zero, and the particle number must be taken as $n = 0$. Clearly, only a restricted range of properties is available to particles which are their own antiparticles; for example, they must be electrically neutral. Examples are the photon and the neutral pion. In the case of the photon, the space and time derivatives of its Hermitian field operators are observable quantities, namely electric and magnetic fields.

7.3 The Dirac Equation and Spin-$\frac{1}{2}$ Particles

7.3.1 The Dirac equation

The problems of negative energies and probabilities encountered in connection with the Klein–Gordon equation evidently have something to do with the fact that this equation involves a second time derivative. Dirac attempted to solve these problems by inventing a new wave equation containing only the first time derivative, which is more closely analogous to the non-relativistic Schrödinger equation. As we shall see, it is not in fact possible to solve the problems in this way, and Dirac's theory also makes sense only as a second-quantized field theory. The Dirac equation is nevertheless of vital importance because it predicts the existence of particles with intrinsic angular momentum or *spin* of magnitude $\hbar/2$.

Such particles are indeed observed, electrons being perhaps the most familiar, and the Dirac theory is the proper Lorentz-covariant means of describing them.

Since special relativity treats time and space on more or less the same footing, an equation that contains only the first time derivative can also contain only first spatial derivatives. The equation must therefore be of the form

$$\left(\mathrm{i}\gamma^{\mu}\partial_{\mu} - m\right)\psi(x) = 0 \tag{7.23}$$

where the four coefficients γ^{μ} are constants. We shall see immediately that these coefficients cannot commute with each other. They must therefore be square matrices rather than simple numbers, so the wavefunction $\psi(x)$ must be a column matrix. This wavefunction must also satisfy the Klein–Gordon equation, which simply expresses the relationship between energy and momentum and, indeed, this should be an automatic consequence of the Dirac equation (7.23). Obviously, we get an equation bearing some resemblance to the Klein–Gordon equation if we act twice with the operator $\left(\mathrm{i}\gamma^{\mu}\partial_{\mu} - m\right)$:

$$\left(\mathrm{i}\gamma^{\mu}\partial_{\mu} - m\right)^2\psi(x) = \left(-\gamma^{\mu}\gamma^{\nu}\partial_{\mu}\partial_{\nu} - 2\mathrm{i}m\gamma^{\mu}\partial_{\mu} + m^2\right)\psi(x) = 0. \tag{7.24}$$

Using the original equation (7.23) and the fact that $\partial_{\mu}\partial_{\nu} = \partial_{\nu}\partial_{\mu}$, we can rewrite this as

$$\left(\tfrac{1}{2}\{\gamma^{\mu}, \gamma^{\nu}\}\partial_{\mu}\partial_{\nu} + m^2\right)\psi(x) = 0. \tag{7.25}$$

In order for this to be the same as the Klein–Gordon equation (7.2), the γ matrices must satisfy the condition

$$\{\gamma^{\mu}, \gamma^{\nu}\} \equiv \gamma^{\mu}\gamma^{\nu} + \gamma^{\nu}\gamma^{\mu} = 2\eta^{\mu\nu} \tag{7.26}$$

where $\eta^{\mu\nu}$ is the (μ, ν) component of the Minkowski-spacetime metric tensor (2.8) and is understood to be multiplied by the unit matrix. A set of matrices that obey this condition is said to form a *Clifford algebra*.

The smallest matrices which can be made to obey the Clifford algebra condition are 4 × 4 matrices, and we shall consider only these. Even so, there are infinitely many representations of the algebra; that is, infinitely many sets of four 4 × 4 matrices that satisfy the condition (7.26). Each representation gives a different, but equivalent, mathematical representation of the same physical situation. For this reason, it is possible to derive all the physical consequences of the theory from the fact that the γ matrices satisfy (7.26). Nevertheless, it is often helpful to have in mind at least one possible set of such matrices. A standard representation is

$$\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \qquad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \tag{7.27}$$

where each entry is itself a 2 × 2 matrix, I being the unit matrix and σ^i the *Pauli matrices*, given by

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma^2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{7.28}$$

Readers who are familiar with the non-relativistic theory of spin-$\frac{1}{2}$ particles, or who have studied appendix B, will recognize these matrices as the wave-mechanical operators that represent the three components of the particle's intrinsic angular momentum. We shall shortly see that this is no coincidence.

7.3.2 Lorentz covariance and spin

As we have discussed in some detail in previous chapters, the equations that express laws of physics are expected to take the same form when referred to any frame of reference. In Minkowski spacetime, it is usually convenient to consider only inertial, Cartesian frames, in which the metric tensor has the simple form (2.8). Because of this restriction, the metric tensor does not appear explicitly in most of our equations, and we expect the form of these equations to remain the same only when we make Lorentz transformations of the form (3.25) (or, more generally, Poincaré transformations, which include spacetime translations). In classical physics, this property of Lorentz covariance is guaranteed if all equations can be expressed in terms of 4-tensors. A Lorentz transformation rearranges the components of a tensor amongst themselves, in such a way that the form of the tensor equations is preserved. In (7.2), we assumed that the wavefunction $\phi(x)$ was the simplest kind of tensor, namely a scalar. The detailed meaning of this is as follows. Suppose that the state of the particle is described by two observers, using sets of coordinates x and x', related by (3.25). The same state will be described by these observers in terms of two wavefunctions $\phi(x)$ and $\phi'(x')$. In general, ϕ and ϕ' are different functions, but if x and x' are the coordinates of the same spacetime point, then $\phi(x) = \phi'(x')$. Since $\partial_\mu \partial^\mu = \partial_{\mu'} \partial^{\mu'}$, each wavefunction also satisfies the Klein–Gordon equation written in its own set of coordinates.

The Dirac wavefunction is a four-component column matrix, so we may expect that, on transforming to a new frame of reference, not only will the components be different functions of the new coordinates, but they will also be rearranged amongst themselves. It turns out that this rearrangement is not the same as those specified by any of the tensor transformation laws (2.19). Although ψ has four components, these do not refer to spacetime directions as do the components of a 4-vector. They actually refer, as we shall see, to different states in which the particle can exist. I shall label these components as ψ_α, where α has the values $1, \ldots, 4$. Thus, ψ is a geometrical object of a kind that we have not previously met. It is called a *spinor*, and its transformation law can be written as

$$\psi'_\alpha(x') = S_{\alpha\beta}(\Lambda)\psi_\beta(x). \tag{7.29}$$

The two sets of coordinates are again related by (3.25). The new transformation matrix S is usually represented as a function of the matrix Λ as I have done here, but it is probably clearer to think of S and Λ as different matrices, which both depend on the same parameters, namely rotation angles and boost velocities such as those in (3.26) and (3.27).

If the Dirac equation is to be covariant, then the transformed wavefunction must satisfy the equation

$$\left(\mathrm{i}\gamma^{\mu'}\partial_{\mu'} - m\right)\psi'(x') = 0. \tag{7.30}$$

The transformed derivative is $\partial_{\mu'} = \Lambda^{\mu}{}_{\mu'}\partial_{\mu}$, and it might appear that the γ matrices should transform as a contravariant 4-vector. This is not correct, though. A constant 4-vector singles out a special direction in spacetime, and the whole point of covariance is that no such special direction exists. The new equation (7.30) is supposed to have the same form as the original equation (7.23), and this means that both observers are entitled to use the same set of γ matrices. Thus, if the old matrices have the numerical values in (7.27) and (7.28), then so do the new ones. But the index μ' indicates that they are associated with the x' coordinate axes. From this requirement, we can work out what the spinor transformation matrix must be. We substitute (7.29) into (7.30), rewrite $\partial_{\mu'}$ in terms of ∂_{μ}, and multiply by S^{-1} to get

$$\left(\mathrm{i}S^{-1}(\Lambda)\gamma^{\mu'}S(\Lambda)\Lambda^{\mu}{}_{\mu'}\partial_{\mu} - m\right)\psi(x) = 0. \tag{7.31}$$

Remembering that $\Lambda^{\mu}{}_{\mu'}\Lambda^{\mu'}{}_{\nu} = \delta^{\mu}_{\nu}$, we see that this is the same as (7.23) provided that

$$S^{-1}(\Lambda)\gamma^{\mu'}S(\Lambda) = \Lambda^{\mu'}{}_{\mu}\gamma^{\mu}. \tag{7.32}$$

Only if a matrix S with this property can be found will the Dirac equation be Lorentz covariant.

It is sufficient to find S for the case of infinitesimal transformations. This will give us the generators of Lorentz transformations, and the matrix for finite transformations can be built up by exponentiation, in just the same way as for spacetime translations. By expanding (3.26) and (3.27) in powers of the rotation angle or boost velocity, we see that Λ can be written as

$$\Lambda^{\mu'}{}_{\mu} = \delta^{\mu'}_{\mu} + \eta^{\mu'\nu'}\omega_{\nu'\mu} + \cdots \tag{7.33}$$

where $\omega_{\nu'\mu}$ is antisymmetric in its two indices, each of its components being proportional to a rotation angle or boost velocity. A general transformation, which is some combination of rotations and boosts, can be written in the same way. Usually, it is meaningless to write symbols like δ and ω with two indices belonging to different coordinate systems. Here it does make sense, because the two sets of coordinates differ only by an infinitesimal amount. The matrix S must be a function of $\omega_{\mu\nu}$ (where it is no longer necessary to distinguish between μ and μ'), and we write its infinitesimal form as

$$S(\Lambda) = I - \frac{\mathrm{i}}{4}\omega_{\mu\nu}\sigma^{\mu\nu} + \cdots. \tag{7.34}$$

In this expression, I is the unit 4×4 matrix and $\sigma^{\mu\nu}$ denotes a set of 4×4 matrices to be constructed in terms of the γ^{μ}. Since $\omega_{\mu\nu}$ is antisymmetric, we can assume that $\sigma^{\mu\nu}$ is also antisymmetric in μ and ν, because a symmetric part would give zero when the implied summations have been carried out. (This antisymmetry means, for example, that $\sigma^{12} = -\sigma^{21}$, but σ^{12} is not necessarily an antisymmetric matrix.) The inverse matrix is $S^{-1} = I + \frac{1}{4}\mathrm{i}\omega_{\mu\nu}\sigma^{\mu\nu} + \cdots$, and if we substitute this together with (7.33) and (7.34) into the condition (7.32), it becomes

$$\left[\gamma^{\lambda}, \sigma^{\mu\nu}\right] = 2\mathrm{i}\left(\eta^{\lambda\mu}\gamma^{\nu} - \eta^{\lambda\nu}\gamma^{\mu}\right). \tag{7.35}$$

Readers may verify using (7.26) that this is satisfied if we identify $\sigma^{\mu\nu}$ as

$$\sigma^{\mu\nu} = \frac{\mathrm{i}}{2}\left[\gamma^{\mu}, \gamma^{\nu}\right]. \tag{7.36}$$

The physical significance of the matrix nature of the Dirac wavefunction can be found by the same method that we used in §5.3 to identify the energy and momentum operators. The momentum operator (5.6) is the generator of space translations, in the sense that it generates a Taylor series like (5.51) when we express a function of the new coordinates $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{a}$ in terms of the old ones. We can carry this idea over to Lorentz transformations, but a slight change of notation will be necessary to distinguish the two sets of coordinates. Just for the purposes of this discussion, I shall replace the notation $x^{\mu'}$ for the new coordinates with $\bar{x}^{\mu}$. Again, this makes sense only because the new coordinate directions differ infinitesimally from the old ones. Consider first a scalar function. Using (7.33), we can write

$$x^{\mu} = \bar{x}^{\mu} - \eta^{\mu\nu}\omega_{\nu\sigma}\bar{x}^{\sigma} + \cdots \tag{7.37}$$

and

$$\phi'(\bar{x}) = \phi(x) = \left(1 - \eta^{\mu\nu}\omega_{\nu\sigma}\bar{x}^{\sigma}\frac{\partial}{\partial\bar{x}^{\mu}} + \cdots\right)\phi(\bar{x}). \tag{7.38}$$

If we take into account the antisymmetry of $\omega_{\mu\nu}$, and use p^{μ} to stand for the wave-mechanical momentum operator $\mathrm{i}\eta^{\mu\nu}\partial_{\nu}$, this can be rewritten as

$$\phi'(x) = \left(1 - \frac{\mathrm{i}}{2}\omega_{\mu\nu}(x^{\mu}p^{\nu} - x^{\nu}p^{\mu}) + \cdots\right)\phi(x). \tag{7.39}$$

This describes the relationship between the functional forms of the old and new wavefunctions, and we can drop the bars over the coordinates, which are now just dummy variables. For Dirac spinors, we must use the transformation law (7.29), and we get an extra term from the matrix S. The result is

$$\psi'(x) = \left(I - \frac{\mathrm{i}}{2}\omega_{\mu\nu}M^{\mu\nu} + \cdots\right)\psi(x) \tag{7.40}$$

with the generators of Lorentz transformations given by

$$M^{\mu\nu} = \tfrac{1}{2}\sigma^{\mu\nu} + \left(x^{\mu}p^{\nu} - x^{\nu}p^{\mu}\right). \tag{7.41}$$

Since these generators are antisymmetric in μ and ν, only six of them are independent. It is useful to divide them into two groups of three, defined by

$$K^i = M^{0i} \qquad J^i = \tfrac{1}{2}\epsilon^{ijk}M^{jk} \tag{7.42}$$

where ϵ^{ijk} is the three-dimensional Levi-Civita tensor, equal to 1 if (i, j, k) is an even permutation of $(1, 2, 3)$, -1 for an odd permutation and zero if any two indices are equal. Thus $J^1 = M^{23}$, $J^2 = M^{31}$ and $J^3 = M^{12}$. The quantity K^i is the generator of boosts along the ith spatial axis, and J^i is the generator of rotations about the ith axis. It is worth noting that a rotation 'about the z axis', say, is more properly described as a rotation in the x-y plane. That is, it rearranges the x and y coordinates, leaving z and t unchanged. The totally antisymmetric tensor ϵ^{ijk} exists only in three spatial dimensions, so in other numbers of dimensions (should we want to consider them) the J^i could not be defined.

The J^i can be written in three-dimensional notation as

$$J^i = \tfrac{1}{2}\begin{pmatrix} \sigma^i & 0 \\ 0 & \sigma^i \end{pmatrix} + (\boldsymbol{r} \times \boldsymbol{p})^i I \tag{7.43}$$

as long as the representation (7.27) is used for the γ matrices. The second term is, of course, the wave-mechanical operator representing the 'orbital' angular momentum associated with the motion of the particle, and we therefore interpret the first term as representing an intrinsic angular momentum or *spin*, which is independent of the orbital motion. Although the Dirac equation is a relativistic one, the existence of particles with spin need not be thought of as a relativistic effect. The generators (7.43) are concerned only with spatial rotations and can be used perfectly well in a non-relativistic theory, as is reviewed in appendix B. In the non-relativistic setting, the independent spin polarization states are specified by the eigenvalue of one spin component, conventionally σ^3, which implies choosing a particular direction in space as the 'spin quantization axis'. In a relativistic theory, this has no Lorentz-covariant meaning, because a Lorentz boost mixes spatial and temporal directions.

A covariant description of spin polarization can be given in terms of the *Pauli–Lubanski 4-vector*, defined by

$$W_\mu = \tfrac{1}{2}\epsilon_{\mu\nu\lambda\sigma}M^{\nu\lambda}p^\sigma \tag{7.44}$$

where $\epsilon_{\mu\nu\lambda\sigma}$ is the four-dimensional Levi-Civita tensor (see appendix A). Since ϵ is totally antisymmetric, we have $\epsilon_{\mu\nu\lambda\sigma}p^\lambda p^\sigma = \epsilon_{\mu\nu\lambda\sigma}p^\nu p^\sigma = 0$, so the $(x^\nu p^\lambda - x^\lambda p^\nu)$ part of $M^{\nu\lambda}$ makes no contribution to W_μ. In terms of the 3-vectors $\boldsymbol{p}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{K}$, where $\boldsymbol{\Sigma}$ is the spin part of $\boldsymbol{J}$, the components of W^μ are

$$W^0 = \boldsymbol{\Sigma} \cdot \boldsymbol{p} \tag{7.45}$$

$$W^i = \Sigma^i p^0 + (\boldsymbol{K} \times \boldsymbol{p})^i. \tag{7.46}$$

The Lorentz-invariant quantity $W^2 = W_\mu W^\mu$ can be evaluated by choosing any convenient frame of reference. If we imagine W^2 to act on a momentum eigenfunction, we can replace p^μ with the corresponding eigenvalue k^μ. By choosing the rest frame of the particle, where $k^\mu = (m, \mathbf{0})$, we find

$$W^2 = -m^2\mathbf{\Sigma}^2 \tag{7.47}$$

and, according to the general theory of angular momentum in quantum mechanics, this should equal $-m^2 s(s+1)$ for a particle of spin s. Thus, a scalar wavefunction with $\mathbf{\Sigma} = 0$ represents a spin-0 particle. For a Dirac spinor, $\mathbf{\Sigma}$ is the matrix in (7.43) and $\mathbf{\Sigma}^2$ is $\frac{3}{4}$ times the unit matrix, so the spinor represents spin-$\frac{1}{2}$ particles.

7.3.3 Some properties of the γ matrices

A number of useful properties of the γ matrices follow from the Dirac equation and the Clifford algebra condition. I shall list several of them, leaving details of their proofs to readers. First, it follows from (7.26) that

$$(\gamma^0)^2 = I \quad \text{and} \quad (\gamma^i)^2 = -I \tag{7.48}$$

for $i = 1, 2$ or 3. If we multiply the Dirac equation (7.23) by γ^0, we get a relativistic Schrödinger equation

$$\mathrm{i}\frac{\partial\psi}{\partial t} = H\psi = \left(-\mathrm{i}\gamma^0\gamma^i\partial_i + m\gamma^0\right)\psi. \tag{7.49}$$

The Hamiltonian H must be Hermitian, and from this it follows that γ^0 is Hermitian and the γ^i are anti-Hermitian:

$$\gamma^{0\dagger} = \gamma^0 \quad \text{and} \quad \gamma^{i\dagger} = -\gamma^i. \tag{7.50}$$

According to (7.26), γ^0 anticommutes with each γ^i, so for $\mu = 0, \ldots, 3$, we can write

$$\gamma^{\mu\dagger} = \gamma^0\gamma^\mu\gamma^0. \tag{7.51}$$

The matrix γ^5 is defined by

$$\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3 = \frac{\mathrm{i}}{4!}\epsilon_{\mu\nu\lambda\sigma}\gamma^\mu\gamma^\nu\gamma^\lambda\gamma^\sigma. \tag{7.52}$$

It has the properties

$$(\gamma^5)^2 = I \tag{7.53}$$
$$\gamma^\mu\gamma^5 = -\gamma^5\gamma^\mu \qquad \text{for any } \mu. \tag{7.54}$$

Although the four matrices γ^μ do not constitute a 4-vector in the ordinary sense, it is often necessary to form contractions as if they did. A useful abbreviation is the 'slash' notation

$$\not{a} \equiv \gamma^\mu a_\mu \tag{7.55}$$

where a_μ is any 4-vector. In this notation, the Dirac equation (7.23) takes the form

$$(\mathrm{i}\not{\partial} - m)\psi(x) = 0. \tag{7.56}$$

The Pauli–Lubanski vector (7.44) can be written, for Dirac spinors, as

$$W_\mu = -\tfrac{1}{4}\left[\gamma_\mu, \not{p}\right]\gamma^5 \tag{7.57}$$

as readers are invited to prove in exercise 7.6.

7.3.4 Conjugate wavefunction and the Dirac action

The adjoint of the Dirac equation (7.23) is

$$\psi^\dagger(x)\left(\mathrm{i}\gamma^{\mu\dagger}\overleftarrow{\partial}_\mu + m\right) = 0 \tag{7.58}$$

where $\overleftarrow{\partial}_\mu$ indicates differentiation of the function on its left. This notation is useful in conjunction with the multiplication of the row matrix $\psi^\dagger$ by a γ matrix on its right. If we multiply this equation from the right by γ^0 and use (7.51), we get

$$\bar{\psi}(x)\left(\mathrm{i}\overleftarrow{\not{\partial}} + m\right) = 0 \tag{7.59}$$

where the conjugate wavefunction is defined by

$$\bar{\psi}(x) = \psi^\dagger(x)\gamma^0. \tag{7.60}$$

It is simple to verify that the two equations (7.56) and (7.59) can be derived as Euler–Lagrange equations from the action

$$S = \int \mathrm{d}^4x\, \bar{\psi}\left(\mathrm{i}\not{\partial} - m\right)\psi \tag{7.61}$$

by treating ψ and $\bar{\psi}$ as independent variables.

7.3.5 Probability current and bilinear covariants

As in the case of scalar wavefunctions, we would like to identify a 4-vector probability current density which is conserved; that is, it satisfies the equation of continuity. The quantity

$$j^\mu(x) = \bar{\psi}(x)\gamma^\mu\psi(x) \tag{7.62}$$

is easily shown, using the Dirac equation and its adjoint, to be conserved. The component $j^0 = \psi^\dagger\psi$, which we would like to identify as the conserved probability density, is positive definite. This would appear to be an advantage, compared with the negative probabilities encountered for the scalar wavefunction, but it will turn out that this is, in a sense, illusory. Since the γ^μ are not themselves the components of a 4-vector, we must show that (7.62) *is* a 4-vector. To do this, we need a property of the transformation matrix $S(\Lambda)$ which, on exponentiating (7.34), is seen to be of the form $S(\Lambda) = \exp(-\mathrm{i}\omega_{\mu\nu}\sigma^{\mu\nu}/4)$. Because of the relation (7.51), we have

$$S^{\dagger}(\Lambda) = \gamma^0 S^{-1}(\Lambda)\gamma^0. \tag{7.63}$$

Using this and the defining property (7.32), we can write the current density in a new frame of reference as

$$\begin{aligned} j^{\mu'}(x') &= \psi'^{\dagger}(x')\gamma^0\gamma^{\mu'}\psi'(x') \\ &= \psi^{\dagger}(x)S^{\dagger}(\Lambda)\gamma^0\gamma^{\mu'}S(\Lambda)\psi(x) \\ &= \bar{\psi}(x)S^{-1}(\Lambda)\gamma^{\mu'}S(\Lambda)\psi(x) \\ &= \Lambda^{\mu'}{}_{\mu}j^{\mu}(x) \end{aligned}$$

so j^{μ} does indeed transform as a 4-vector. Note that the presence of $\bar{\psi}$ rather than $\psi^{\dagger}$ is essential to this proof.

A number of other tensors can be constructed in the same way. To understand how these are classified, it is necessary to consider a wider class of Lorentz transformations than we have so far. The representative transformation matrices (3.26) and (3.27) each have $\Lambda^0{}_0 \geq 1$ and $\det(\Lambda) = +1$. Such transformations are called *proper* Lorentz transformations. Examples of 'improper' transformations are *time reversal* $t' = -t$ and *parity* or spatial reflection $\boldsymbol{x}' = -\boldsymbol{x}$. Each of these has $\det(\Lambda) = -1$. Several important tensor-like quantities have transformation laws similar to (2.19), except that the right-hand side is multiplied by $\det(\Lambda)$. These are called *pseudotensors*. Three-dimensional examples are provided by the cross products $\boldsymbol{a} \times \boldsymbol{b}$ of any two vectors, which are called *axial vectors*. Each vector changes sign under parity, but the product does not change sign. (More generally, a quantity whose transformation law contains a factor $[\det(\Lambda)]^n$ is a *tensor density* of weight n.)

The so-called *bilinear covariants* are products of the form $\bar{\psi}\Gamma\psi$, where Γ is a 4×4 matrix. Any 4×4 matrix can be written as a linear combination of 16 linearly independent ones. Such a set is provided by the matrices I, γ^5, γ^{μ}, $\gamma^{\mu}\gamma^5$ and $\sigma^{\mu\nu}$, which have the advantage of giving rise to tensors or pseudotensors. The names given to these objects and their transformation properties are

scalar: $S(x) = \bar{\psi}(x)\psi(x)$	$S'(x') = S(x)$
pseudoscalar: $P(x) = \bar{\psi}(x)\gamma^5\psi(x)$	$P'(x') = \det(\Lambda)P(x)$
vector: $V^{\mu}(x) = \bar{\psi}(x)\gamma^{\mu}\psi(x)$	$V^{\mu'}(x') = \Lambda^{\mu'}{}_{\mu}V^{\mu}(x)$
axial vector: $A^{\mu}(x) = \bar{\psi}(x)\gamma^{\mu}\gamma^5\psi(x)$	$A^{\mu'}(x') = \det(\Lambda)\Lambda^{\mu'}{}_{\mu}A^{\mu}(x)$
tensor: $T^{\mu\nu}(x) = \bar{\psi}(x)\sigma^{\mu\nu}\psi(x)$	$T^{\mu'\nu'}(x') = \Lambda^{\mu'}{}_{\mu}\Lambda^{\nu'}{}_{\nu}T^{\mu\nu}(x)$.

The vector covariant is, of course, the same as (7.62), and the proofs of all the transformation properties are similar to that given above.

7.3.6 Plane-wave solutions

As in the non-relativistic theory, a complete set of plane-wave solutions to the Dirac equation is labelled by the momentum $\boldsymbol{k}$ and a spin component $s = \pm\frac{1}{2}$ along a chosen quantization axis. A covariant description of the spin polarization of a massive particle can be given as follows. In the rest frame, where $k^\mu = (m, \mathbf{0})$, choose a unit 3-vector $\boldsymbol{n}$ as the quantization axis. In a frame in which the momentum is $(k^0, \boldsymbol{k})$, the object

$$n^\mu = \left(\frac{\boldsymbol{k}\cdot\boldsymbol{n}}{m},\ \boldsymbol{n} + \frac{(\boldsymbol{k}\cdot\boldsymbol{n})}{m(m+k^0)}\,\boldsymbol{k}\right) \tag{7.64}$$

is a 4-vector, with $n_\mu n^\mu = -1$ and $k_\mu n^\mu = 0$. The quantity $W \cdot n = W_\mu n^\mu$ is Lorentz invariant. Its value is most easily calculated in the rest frame and is

$$W \cdot n = -m\boldsymbol{\Sigma}\cdot\boldsymbol{n} \tag{7.65}$$

which is the component of spin along $\boldsymbol{n}$ as measured in the rest frame. A complete set of plane-wave solutions is now given by the simultaneous eigenfunctions of $W \cdot n$ and the momentum operator $\mathrm{i}\partial_\mu$. There are both positive- and negative-energy solutions. Let $k^0 = +\left(\boldsymbol{k}^2 + m^2\right)^{1/2}$. The positive-energy solutions have the form

$$\psi_{k,s}(x) = \mathrm{e}^{-\mathrm{i}k\cdot x}u(k,s) \tag{7.66}$$

where $u(k,s)$ is a column matrix. To satisfy the Dirac equation (7.56), we must have

$$(\not{k} - m)\,u(k,s) = 0 \tag{7.67}$$

and, according to the above definition of spin polarization, $(W \cdot n)u(k,s) = -ms\,u(k,s)$. This means that s is the spin component in the direction $\boldsymbol{n}$ *that would be measured by an observer in the particle's rest frame*, even when $u(k,s)$ describes the state as observed in some other frame. If we do consider the rest frame, and choose $\boldsymbol{n} = (0,0,1)$, then with the standard representation (7.27) for the γ matrices we find

$$u(k,\tfrac{1}{2}) = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \qquad \text{and} \qquad u(k,-\tfrac{1}{2}) = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}. \tag{7.68}$$

Corresponding to each positive-energy solution there is a negative-energy solution

$$\psi^{\mathrm{c}}_{k,s}(x) = \mathrm{e}^{\mathrm{i}k\cdot x}v(k,s) \tag{7.69}$$

where the negative-energy spinor $v(k,s)$ satisfies

$$(\not{k} + m)v(k,s) = 0. \tag{7.70}$$

As in the scalar theory, it will be necessary to reinterpret these negative-energy solutions in terms of antiparticles. In the scalar case, the negative-energy solution is the complex conjugate of a positive-energy antiparticle wavefunction. Here, (7.69) is the *charge conjugate* of a positive-energy antiparticle wavefunction. The operation of charge conjugation, denoted by the superscript c in (7.69), relates particle and antiparticle states. It involves both complex conjugation and a rearrangement of spinor components. To find the positive-energy solution of which (7.69) is the conjugate, we define

$$\psi^{\mathrm{c}}_{k,s}(x) = \mathcal{C}\psi^{*}_{k,s}(x) \tag{7.71}$$

where $\mathcal{C}$ is a matrix to be found. The spinor $v(k,s) = \mathcal{C}u^*(k,s)$ must satisfy (7.70), given that $u(k,s)$ satisfies (7.67). Taking the complex conjugate of (7.67) and multiplying by $\mathcal{C}$, we find that this will be so provided that

$$\mathcal{C}\gamma^{\mu *}\mathcal{C}^{-1} = -\gamma^{\mu}. \tag{7.72}$$

This is usually expressed differently, by observing that $\gamma^{\mu *}$ is the transpose (denoted by $^{\mathrm{T}}$) of $\gamma^{\mu\dagger}$. Then by using (7.48) and (7.51), we can express $\mathcal{C}$ as $\mathcal{C} = C\,\gamma^{0\mathrm{T}}$, where the charge conjugation matrix C has the property

$$C\,\gamma^{\mu\mathrm{T}}C^{-1} = -\gamma^{\mu}. \tag{7.73}$$

This relation does not define C uniquely; the usual choice of a matrix that works, within the standard representation of the γ matrices, is $C = \mathrm{i}\gamma^2\gamma^0$. The charge conjugate spinors corresponding to (7.68) are

$$v(k,\tfrac{1}{2}) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \qquad \text{and} \qquad v(k,-\tfrac{1}{2}) = -\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}. \tag{7.74}$$

Some further properties of charge conjugation are explored in the exercises, as is the construction of plane-wave solution in frames other than the rest frame.

7.3.7 Massless spin-$\frac{1}{2}$ particles

A spin-$\frac{1}{2}$ particle whose mass is zero satisfies the Dirac equation $\mathrm{i}\not{\partial}\psi = 0$. Whether such particles exist in nature is uncertain. Neutrinos are spin-$\frac{1}{2}$ particles and their masses are too small to be measured directly, but there is (at the time of writing) some indirect evidence to suggest that some at least of the known neutrino species have non-zero masses. Be that as it may, solutions of the massless Dirac equation play an important role in several theories that we shall examine later on. A massless particle travels with the speed of light and therefore has no rest frame, so the polarization vector (7.64) cannot be defined. Instead, spin states can be classified according to *helicity*, which is the component of spin parallel to the 3-vector momentum $\boldsymbol{k}$:

$$h = \boldsymbol{\Sigma} \cdot \boldsymbol{k}/|\boldsymbol{k}|. \tag{7.75}$$

The Pauli–Lubanski vector can be expressed as $W^\mu = -\frac{1}{2}\gamma^5\left[\gamma^\mu \not{k} - k^\mu\right]$ so, for a wavefunction satisfying the massless Dirac equation, we have $W^\mu \psi = \frac{1}{2}\gamma^5 k^\mu \psi$. Thus, a plane-wave solution with a definite momentum k^μ will also be an eigenfunction of W^μ if it is an eigenfunction of γ^5. Any wavefunction can be decomposed as $\psi = \psi_{\mathrm{R}} + \psi_{\mathrm{L}}$, where

$$\psi_{\mathrm{R}} = \tfrac{1}{2}(1+\gamma^5)\psi \qquad \psi_{\mathrm{L}} = \tfrac{1}{2}(1-\gamma^5)\psi. \tag{7.76}$$

(Here and in other similar contexts, I follow the custom of using ‘1’ to denote the unit matrix.) Since $(\gamma^5)^2 = 1$, these two components are eigenfunctions of γ^5, with eigenvalues $+1$ and -1 respectively. If ψ is a plane wave, with momentum eigenvalue k^μ, they are eigenfunctions of W^μ with eigenvalues $\pm\frac{1}{2}k^\mu$. In particular, they are eigenfunctions of W^0 with eigenvalues $\pm\frac{1}{2}k^0$. Since $W^0 = \boldsymbol{\Sigma} \cdot \boldsymbol{k}$ and $k^0 = |\boldsymbol{k}|$ for a massless particle, we find that the component ψ_{R} has helicity $h = +\frac{1}{2}$ while ψ_{L} has helicity $h = -\frac{1}{2}$. If we picture a positive-helicity particle as a small spinning sphere, whose angular momentum is parallel to $\boldsymbol{k}$, then the fingers of a right hand whose thumb is extended in the direction of $\boldsymbol{k}$ would curl in the direction of the sphere’s rotation. In this sense, the component ψ_{R} is said to be right-handed, while ψ_{L} is left-handed. For any spinor, these components are called the *chiral projections* and in this context γ^5 is the *chirality* or ‘handedness’ operator. However, it is only for massless particles that these chiral projections have definite helicities.

7.4 Spinor Field Theory

Although the Dirac equation appears to lead to a positive definite probability density, and thus to solve one of the problems that we encountered in interpreting solutions of the Klein–Gordon equation, it nevertheless has negative-energy solutions, as we have seen. In order to interpret these in terms of antiparticles, we must again resort to second quantization. If we write out matrix multiplications explicitly, the action (7.61) is

$$S = \int \mathrm{d}^4x\, \bar{\psi}_i \left(\mathrm{i}\gamma^\mu_{ij}\partial_\mu - m\delta_{ij}\right)\psi_j. \tag{7.77}$$

The momentum conjugate to ψ_i is

$$\Pi_i = \frac{\delta S}{\delta(\partial_0 \psi_i)} = \mathrm{i}\bar{\psi}_j \gamma^0_{ji} = \mathrm{i}\psi_i^\dagger \tag{7.78}$$

which is the same as (6.26) for the non-relativistic Schrödinger theory. When ψ satisfies the Dirac equation, the action is zero, and in that case the Hamiltonian is

$$H = \int \mathrm{d}^3x\, \Pi_i(\boldsymbol{x},t)\dot{\psi}_i(\boldsymbol{x},t) = \int \mathrm{d}^3x\, \bar{\psi}(\boldsymbol{x},t)\mathrm{i}\gamma^0\partial_0\psi(\boldsymbol{x},t). \tag{7.79}$$

In accordance with our earlier procedure, we replace the wavefunction with a field operator. This may be expanded in terms of plane-wave solutions as

$$\hat{\psi}(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})} \sum_s \left[\hat{b}(\boldsymbol{k}, s)\mathrm{e}^{-\mathrm{i}k\cdot x} u(k, s) + \hat{d}^\dagger(\boldsymbol{k}, s)\mathrm{e}^{\mathrm{i}k\cdot x} v(k, s)\right] \tag{7.80}$$

in which $k^0 = (\boldsymbol{k}^2 + m^2)^{1/2}$. The operator $\hat{b}(\boldsymbol{k}, s)$ is to be interpreted as the annihilation operator for a particle of 3-momentum $\boldsymbol{k}$ and spin polarization s, and $\hat{d}^\dagger(\boldsymbol{k}, s)$ as the creation operator for an antiparticle. It is possible to normalize $u(k, s)$ and $v(k, s)$ in such a way that

$$\bar{u}(k, s)\gamma^\mu u(k, s') = \bar{v}(k, s)\gamma^\mu v(k, s') = 2k^\mu \delta_{ss'} \tag{7.81}$$

$$\bar{u}(k, s)\gamma^0 v(\bar{k}, s') = \bar{v}(k, s)\gamma^0 u(\bar{k}, s') = 0 \tag{7.82}$$

where $\bar{k}^\mu = (k^0, -\boldsymbol{k})$ (see exercise 7.4), and this leads to the same covariant normalization for the particle states as we had for spin-0 particles. In particular, the creation and annihilation operators can be expressed in terms of $\hat{\psi}$ through

$$\hat{b}(\boldsymbol{k}, s) = \int \mathrm{d}^3x\, \mathrm{e}^{\mathrm{i}k\cdot x} \bar{u}(k, s)\gamma^0 \hat{\psi}(x) \tag{7.83}$$

$$\hat{d}^\dagger(\boldsymbol{k}, s) = \int \mathrm{d}^3x\, \mathrm{e}^{-\mathrm{i}k\cdot x} \bar{v}(k, s)\gamma^0 \hat{\psi}(x) \tag{7.84}$$

which correspond to (7.12) and (7.13) in the scalar theory.

In terms of the creation and annihilation operators, the Hamiltonian reads

$$\hat{H} = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})} \sum_s \omega(\boldsymbol{k}) \left[\hat{b}^\dagger(\boldsymbol{k}, s)\hat{b}(\boldsymbol{k}, s) - \hat{d}(\boldsymbol{k}, s)\hat{d}^\dagger(\boldsymbol{k}, s)\right]. \tag{7.85}$$

If we were to assume commutation relations similar to (7.17), it may be seen that the antiparticles would contribute negative energies. Other undesirable consequences would also follow. For example, causality would be violated, in the sense that operators representing observable quantities in regions of spacetime at space-like separations would fail to commute. Thus, events in these regions, which cannot communicate via signals travelling at speeds less than or equal to that of light, would not be independent as they ought to be. It is these inconsistencies that give rise to the *spin-statistics theorem* mentioned in §6.1. They can be removed if we assume instead the *anticommutation relations*

$$\{\hat{b}(\boldsymbol{k}, s), \hat{b}^\dagger(\boldsymbol{k}', s')\} = \{\hat{d}(\boldsymbol{k}, s), \hat{d}^\dagger(\boldsymbol{k}', s')\} = (2\pi)^3 2\omega(\boldsymbol{k})\delta_{ss'}\delta(\boldsymbol{k} - \boldsymbol{k}') \tag{7.86}$$

with all other anticommutators equal to zero. The antiparticle term in (7.85) then changes sign when we reverse the order of the operators, and we also get an infinite constant, as in (7.21). Removing the constant is again equivalent to normal ordering, provided that the definition of normal ordering is amended to include a change of sign whenever two fermionic operators are interchanged. The relations (7.86) imply equal-time anticommutation relations for the field components, which are

$$\{\hat{\psi}_i(\boldsymbol{x}, t), \hat{\Pi}_j(\boldsymbol{x}', t)\} = \mathrm{i}\delta_{ij}\delta(\boldsymbol{x} - \boldsymbol{x}') \tag{7.87}$$

the anticommutator of two field components or two momentum components being zero.

When anticommutation is taken into account, the number operator for spin-$\frac{1}{2}$ fermions is found to be

$$\begin{aligned} \hat{N} &= \int \mathrm{d}^3x \, :\hat{\bar{\psi}}(\boldsymbol{x}, t)\gamma^0\hat{\psi}(\boldsymbol{x}, t): \\ &= \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})} \sum_s \left[\hat{b}^\dagger(\boldsymbol{k}, s)\hat{b}(\boldsymbol{k}, s) - \hat{d}^\dagger(\boldsymbol{k}, s)\hat{d}(\boldsymbol{k}, s)\right]. \end{aligned} \tag{7.88}$$

This counts the (number of particles − number of antiparticles), which is the desired result. It can, of course, take both positive and negative values, which is ironic, since the positive definite probability density appeared at first to be a success of the Dirac equation. We see, indeed, that at the level of first quantization, the Dirac theory cannot be quite correct. To allow for the antiparticle interpretation, it ought to be possible for $j^0(x)$ to have negative values. There is, in fact, a modification that will do this. Let us consider the plane-wave expansion (7.80) to apply to a wavefunction, the coefficients $b(\boldsymbol{k}, s)$ and $d^*(\boldsymbol{k}, s)$ being numbers rather than operators. For consistency with the anticommutation of the corresponding operators, these should be regarded as *anticommuting numbers*. This means that $b(\boldsymbol{k}, s)b(\boldsymbol{k}', s') = -b(\boldsymbol{k}', s')b(\boldsymbol{k}, s)$ and similarly for any product of bs, ds and their complex conjugates. In particular, the product of an anticommuting number with itself is zero. However, any anticommuting number still commutes with an ordinary commuting (or c-) number. Such anticommuting numbers are said to form a *Grassmann algebra* (see appendix A). The Dirac wavefunction itself is therefore also an anticommuting Grassmann number. For many purposes, we deal only with equations which, like the Dirac equation itself, are linear in the wavefunction, so the anticommutation has no effect. None of the results derived in previous sections are changed. However, certain properties of the bilinear covariants do depend on whether the wavefunction is taken to be commuting or anticommuting, and these will be consistent with corresponding properties of the second-quantized operators only if anticommuting wavefunctions are used. The Hamiltonian and current density are cases in point.

7.5 Weyl and Majorana Spinors

We can now see in detail the physical meaning of the four components of a Dirac spinor. The four degrees of freedom correspond to four single-particle states: a particle and an antiparticle state, each of which can have either of two independent spin polarizations. A question that turns out to be worth asking is this: is it necessarily true that a spin-$\frac{1}{2}$ particle species has all four of these states available to it? The considerations involved in addressing this question are especially important in the case of massless particles, and most of our discussion will focus on these. The algebra that is needed becomes particularly straightforward if we choose a different representation of the γ matrices from the standard set (7.27) that we have used up to now. The four matrices

$$\gamma^0 = \begin{pmatrix} 0 & -I \\ -I & 0 \end{pmatrix} \qquad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} \tag{7.89}$$

satisfy the Clifford algebra condition (7.26) and constitute the *Weyl* or *chiral* representation. In this representation, the matrices γ^5, C and $\mathcal{C}$ that we defined earlier are given by

$$\gamma^5 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \qquad C = \begin{pmatrix} -\epsilon & 0 \\ 0 & \epsilon \end{pmatrix} \qquad \mathcal{C} = \begin{pmatrix} 0 & \epsilon \\ -\epsilon & 0 \end{pmatrix} \tag{7.90}$$

where ϵ is the 2×2 matrix

$$\epsilon = \mathrm{i}\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{7.91}$$

which has the properties $\epsilon\sigma^{i*}\epsilon = \sigma^i$ and $\epsilon^2 = -I$. The operators that appear in the equations (7.67) and (7.70) for positive- and negative-energy spinors are

$$\not{k} \mp m = -\begin{pmatrix} \pm m\, I & (k^0 I + \boldsymbol{\sigma} \cdot \boldsymbol{k}) \\ (k^0 I - \boldsymbol{\sigma} \cdot \boldsymbol{k}) & \pm m\, I \end{pmatrix}. \tag{7.92}$$

For massless particles, the positive- and negative-energy spinors u and v obey the same equation $\not{k}u = \not{k}v = 0$. In fact, there is no need to distinguish between u and v, so I shall write this single equation as $\not{k}u = 0$. A massless particle has both $m = 0$ and $k^0 = |\boldsymbol{k}|$. (Recall from our discussion in §7.3.6 that k^0 is positive for both positive- and negative-energy solutions, but these two types of solution are distinguished by the sign of the exponential factor in (7.66) and (7.69)). Suppose, for simplicity, that $\boldsymbol{k} = (0, 0, k)$. Then $\not{k}$ can be written explicitly as

$$\not{k} = \begin{pmatrix} 0 & 0 & -2k & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2k & 0 & 0 \end{pmatrix}. \tag{7.93}$$

The equation $\not{k}u = 0$ has only two independent solutions, which are

$$u_{\mathrm{R}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad u_{\mathrm{L}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{7.94}$$

with $\gamma^5 u_\mathrm{R} = +u_\mathrm{R}$ and $\gamma^5 u_\mathrm{L} = -u_\mathrm{L}$. Clearly, when $\boldsymbol{k}$ is in some other direction, there can also be only two independent solutions. They can be written as

$$u_\mathrm{R}(k) = \begin{pmatrix} \chi(\boldsymbol{k}) \\ 0 \end{pmatrix} \qquad u_\mathrm{L}(k) = \begin{pmatrix} 0 \\ -\epsilon\chi^*(\boldsymbol{k}) \end{pmatrix} \tag{7.95}$$

where each entry is a 2-component column matrix and

$$\chi(\boldsymbol{k}) = \frac{|\boldsymbol{k}|}{\sqrt{|\boldsymbol{k}| + k^3}} \left(I + \frac{\boldsymbol{\sigma} \cdot \boldsymbol{k}}{|\boldsymbol{k}|} \right) \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{7.96}$$

With the normalization factor given here, the two solutions have the orthonormality properties

$$\bar{u}_\mathrm{R}(k)\gamma^\mu u_\mathrm{R}(k) = \bar{u}_\mathrm{L}(k)\gamma^\mu u_\mathrm{L}(k) = 2k^\mu \tag{7.97}$$

$$\bar{u}_\mathrm{R}(k)\gamma^\mu u_\mathrm{L}(k) = \bar{u}_\mathrm{L}(k)\gamma^\mu u_\mathrm{R}(k) = 0. \tag{7.98}$$

They are also related by charge conjugation:

$$u_\mathrm{R}^\mathrm{c}(k) = \mathcal{C}u_\mathrm{R}^*(k) = u_\mathrm{L}(k) \qquad u_\mathrm{L}^\mathrm{c}(k) = \mathcal{C}u_\mathrm{L}^*(k) = u_\mathrm{R}(k). \tag{7.99}$$

The general solution to the massless Dirac equation can now be written as $\hat{\psi}(x) = \hat{\psi}_\mathrm{R}(x) + \hat{\psi}_\mathrm{L}(x)$, with

$$\hat{\psi}_\mathrm{R}(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2|\boldsymbol{k}|} \left[\hat{b}_\mathrm{R}(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}k\cdot x} u_\mathrm{R}(k) + \hat{d}_\mathrm{L}^\dagger(\boldsymbol{k})\mathrm{e}^{\mathrm{i}k\cdot x} u_\mathrm{R}(k) \right] \tag{7.100}$$

$$\hat{\psi}_\mathrm{L}(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2|\boldsymbol{k}|} \left[\hat{b}_\mathrm{L}(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}k\cdot x} u_\mathrm{L}(k) + \hat{d}_\mathrm{R}^\dagger(\boldsymbol{k})\mathrm{e}^{\mathrm{i}k\cdot x} u_\mathrm{L}(k) \right]. \tag{7.101}$$

Note carefully that, since the charge conjugate of a right-handed solution is a left-handed solution, the coefficient of the negative-energy term in $\hat{\psi}_\mathrm{R}$ must be interpreted as the creation operator for a left-handed particle, and conversely for $\hat{\psi}_\mathrm{L}$. It is not hard to verify that the Dirac equation can be written in terms of the right- and left-handed components as

$$\mathrm{i}\not{\partial}\hat{\psi}_\mathrm{R} = m\hat{\psi}_\mathrm{L} \qquad \mathrm{i}\not{\partial}\hat{\psi}_\mathrm{L} = m\hat{\psi}_\mathrm{R}. \tag{7.102}$$

In the case of massless particles, these are two independent equations. Correspondingly, the result of exercise 7.9 shows that the action (7.61) can be expressed as

$$S = \int \mathrm{d}^4x \left[\mathrm{i}\bar{\psi}_\mathrm{R}\not{\partial}\psi_\mathrm{R} + \mathrm{i}\bar{\psi}_\mathrm{L}\not{\partial}\psi_\mathrm{L} \right]. \tag{7.103}$$

It is therefore possible to delete, say, ψ_R from our theory entirely or, equivalently, to construct a theory that involves only the left-handed field ψ_L. The spinor in this reduced theory is called a *Weyl spinor*. The theory contains two independent annihilation operators, $\hat{b}_{\rm L}(\boldsymbol{k})$ and $\hat{d}_{\rm R}(\boldsymbol{k})$ together with the creation operators $\hat{b}^\dagger_{\rm L}(\boldsymbol{k})$ and $\hat{d}^\dagger_{\rm R}(\boldsymbol{k})$. In this theory, the particles can exist only in the left-handed state, while the antiparticles exist only in the right-handed state. In the alternative theory, which contains only $\psi_{\rm R}$, the converse would be true. However, these two theories are physically equivalent, since we can rename the particles as antiparticles and *vice versa*. They are also mathematically equivalent, because we can rewrite the theory of $\psi_{\rm R}$ in terms of $\psi^{\rm c}_{\rm R}$, which is a left-handed field.

The theory containing a single massless Weyl spinor thus provides one example of a particle that has available to it only two of the four states in the full Dirac theory. A second example is provided by a spin-$\frac{1}{2}$ particle which is its own antiparticle. The field operator $\psi_{\rm M}(x)$ for such a particle must obey $\psi^{\rm c}_{\rm M}(x) = \psi_{\rm M}(x)$ and is called a *Majorana spinor*. In the case of a massless particle, the field can be written in terms of creation and annihilation operators $\hat{\beta}_{\rm R}$, $\hat{\beta}^\dagger_{\rm R}$, $\hat{\beta}_{\rm L}$, $\hat{\beta}^\dagger_{\rm L}$ for the right- and left-handed states available to it:

$$\begin{aligned}\hat{\psi}_{\rm M}(x) = \int \frac{{\rm d}^3k}{(2\pi)^3 2|\boldsymbol{k}|}\Big[&{\rm e}^{-{\rm i}k\cdot x}\Big(u_{\rm R}(k)\hat{\beta}_{\rm R}(\boldsymbol{k}) + u_{\rm L}(k)\hat{\beta}_{\rm L}(\boldsymbol{k})\Big)\\ &+ {\rm e}^{{\rm i}k\cdot x}\Big(u_{\rm R}(k)\hat{\beta}^\dagger_{\rm L}(\boldsymbol{k}) + u_{\rm L}(k)\hat{\beta}^\dagger_{\rm R}(\boldsymbol{k})\Big)\Big].\end{aligned} \tag{7.104}$$

The Dirac equation for a Majorana spinor can be obtained as the Euler–Lagrange equation associated with the action

$$S = \int {\rm d}^4x\, \tfrac{1}{2}{\rm i}\bar{\psi}_{\rm M}\partial\!\!\!/\psi_{\rm M}. \tag{7.105}$$

The factor of $\frac{1}{2}$ is necessary to maintain the anticommutation relations (7.87) for the field and its conjugate momentum, given that we have a reduced number of creation and annihilation operators with the anticommutators

$$\{\hat{\beta}_{\rm A}(\boldsymbol{k}), \hat{\beta}^\dagger_{\rm B}(\boldsymbol{k}')\} = (2\pi)^3 2|\boldsymbol{k}|\delta_{\rm AB}\delta(\boldsymbol{k} - \boldsymbol{k}') \tag{7.106}$$

where the indices A and B have the values R or L.

For non-interacting, massless particles, the theories containing a single Weyl spinor or a Majorana spinor are actually equivalent. Mathematically, we can use the left-handed field (7.101), for example, to build a Majorana spinor

$$\begin{aligned}\hat{\phi}_{\rm M}(x) &= \psi_{\rm L}(x) + \psi^{\rm c}_{\rm L}(x)\\ &= \int \frac{{\rm d}^3k}{(2\pi)^3 2|\boldsymbol{k}|}\Big[{\rm e}^{-{\rm i}k\cdot x}\Big(u_{\rm R}(k)\hat{d}_{\rm R}(\boldsymbol{k}) + u_{\rm L}(k)\hat{b}_{\rm L}(\boldsymbol{k})\Big)\\ &\qquad + {\rm e}^{{\rm i}k\cdot x}\Big(u_{\rm R}(k)\hat{b}^\dagger_{\rm L}(\boldsymbol{k}) + u_{\rm L}(k)\hat{d}^\dagger_{\rm R}(\boldsymbol{k})\Big)\Big]\end{aligned} \tag{7.107}$$

and it is possible to show that the action for this field is

$$S = \int {\rm d}^4x\, \tfrac{1}{2}{\rm i}\bar{\phi}_{\rm M}\partial\!\!\!/\phi_{\rm M} = \int {\rm d}^4x\, {\rm i}\bar{\psi}_{\rm L}\partial\!\!\!/\psi_{\rm L}. \tag{7.108}$$

Moreover, the operators $(\hat{b}_{\rm L}, \hat{b}^\dagger_{\rm L}, \hat{d}_{\rm R}, \hat{d}^\dagger_{\rm R})$ have exactly the same anticommutation relations as $(\hat{\beta}_{\rm L}, \hat{\beta}^\dagger_{\rm L}, \hat{\beta}_{\rm R}, \hat{\beta}^\dagger_{\rm R})$, so the difference between these two theories is purely a matter of notation. From a physical point of view, the available states in each case are a left-handed particle and a right-handed antiparticle. According to the Weyl description, the particle and antiparticle are distinct, while according to the Majorana description they are the same particle. This might seem to be a genuine physical difference. However, the difference is physically undetectable so long as the particles do not interact. If the particles do interact, then the nature of the interaction will tell us which description is appropriate. For example, it is possible to construct an idealized theory of massless electrons, in which electrons are always left-handed and positrons are always right-handed. If these electrons interact via electromagnetic fields, then the extra term in the S needed to account for this interaction (to be discussed in chapter 8) cannot be built from a Majorana spinor. A Weyl spinor is needed to describe the electromagnetic interaction and, of course, electrons and positrons will turn out to be different particles, since they have opposite charges.

7.6 Particles of Spin 1 and 2

In later chapters, we shall encounter fundamental spin-1 particles (photons, which are massless and the $W^\pm$ and Z^0 particles, which are massive). In a quantum theory of gravity, there ought also to be gravitons, which turn out to have spin 2, although these particles have not (at the time of writing) been detected experimentally. All the theories involving these particles give rise to special technical questions, which I shall discuss in due course. In this section, we take a preliminary look at the wave equations that describe such particles in the absence of interactions, and investigate how they can be interpreted in terms of spin.

7.6.1 Photons and massive spin-1 particles

In the absence of charged particles, Maxwell's equations (3.52) can be written in terms of the 4-vector potential A_μ as

$$\Box A_\mu - \partial_\mu(\partial_\nu A^\nu) = 0. \tag{7.109}$$

A modification of this equation, called the *Proca equation*

$$\Box A_\mu + m^2 A_\mu - \partial_\mu(\partial_\nu A^\nu) = 0 \tag{7.110}$$

describes particles of mass m. In fact, if we act with ∂^μ on (7.110), the first and last terms cancel and the remaining equation tells us that

$$\partial_\mu A^\mu = 0. \tag{7.111}$$

Using this result, (7.110) becomes just the Klein–Gordon equation $(\Box + m^2)A_\mu = 0$. In the Maxwell theory, we can use the property of gauge invariance to impose the condition (7.111) on the solutions of (7.109) also. Thus, if we make a gauge transformation (3.60) and choose $\theta(x)$ to be a solution of $\Box\theta = \partial_\mu A^\mu$, the new vector potential $A'_\mu(x)$ obeys (7.111), which in this context is called the *Lorentz gauge condition*. To put this another way, any solution of the Maxwell wave equation (7.109) can be written as $A_\mu(x) = A_\mu^{(\mathrm{L})}(x) + \partial_\mu\theta(x)$, where $A_\mu^{(\mathrm{L})}(x)$ obeys the Lorentz condition. The term $\partial_\mu\theta(x)$ has no physical meaning, since it makes no contribution to the electric and magnetic fields and therefore, according to exercise 3.6, makes no contribution to the energy either, so it can be discarded.

The spin of particles described by a 4-vector wavefunction or field operator can be determined by the same method that we used for Dirac spinors. Let us assemble the four components of A_μ into a column matrix A. Under a Lorentz transformation, we have $A'(x') = \Lambda A(x)$, which is analogous to (7.29), the matrix $S(\Lambda)$ now being just Λ itself. In this matrix language, the generators of Lorentz transformations which appear in the Pauli–Lubanski vector (7.45) and (7.46) are given for 4-vector fields by

$$\Sigma^1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \end{pmatrix} \quad \Sigma^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 & 0 \\ 0 & -\mathrm{i} & 0 & 0 \end{pmatrix} \quad \Sigma^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} & 0 \\ 0 & \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.112}$$

$$K^1 = \begin{pmatrix} 0 & \mathrm{i} & 0 & 0 \\ \mathrm{i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad K^2 = \begin{pmatrix} 0 & 0 & \mathrm{i} & 0 \\ 0 & 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad K^3 = \begin{pmatrix} 0 & 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 & 0 \end{pmatrix}. \tag{7.113}$$

As before, we consider plane-wave solutions, which in this case have the form

$$A_k^\mu(x) = \epsilon^\mu \mathrm{e}^{-\mathrm{i}k\cdot x} \tag{7.114}$$

and represent the components of the *polarization vector* ϵ^μ as a column matrix. The condition (7.111) implies $k_\mu\epsilon^\mu = 0$. For massive particles, we can use the rest frame, where $k^\mu = (m, \mathbf{0})$, and calculate the square of the Pauli–Lubanski vector, with the result

$$W^2 \equiv W_\mu W^\mu = -m^2\mathbf{\Sigma}^2 = -2m^2 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.115}$$

The factor of $s(s+1) = 2$ indicates that the particles have spin $s = 1$, but the matrix here is not the unit matrix. Taking the spin quantization axis in the x^3 direction as usual, we can find a set of four basis vectors ϵ_λ for the polarization (λ

is a label for these vectors, not a spacetime index), which are eigenvectors of Σ^3:

$$\epsilon_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\\mathrm{i}\\0\end{pmatrix} \qquad \epsilon_0 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \qquad \epsilon_{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\-\mathrm{i}\\0\end{pmatrix} \qquad \epsilon_0' = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}. \tag{7.116}$$

The first three of these, with eigenvalues 1, 0 and -1 respectively correspond to the expected values of the $\mu = 3$ component of spin. The last, ϵ_0', also has eigenvalue 0, but it does not obey the condition (7.111), which in the rest frame becomes $m\epsilon^0 = 0$, and so is not an allowed solution. Acting on a solution built from the first three polarization vectors, the matrix in (7.115) is, in effect, just the unit 3×3 matrix.

The spin polarization of a massless particle such as a photon must again be described in terms of helicity, and again we consider a frame of reference in which $k^\mu = (k, 0, 0, k)$ with $k > 0$. In this frame, the helicity operator $h = W^0/k^0$ is $h = \Sigma^3$, and it has the same eigenvectors (7.116). However, the condition $\partial_\mu A^\mu = 0$ now says that $k_\mu \epsilon^\mu = k(\epsilon^0 - \epsilon^3) = 0$, or $\epsilon^0 = \epsilon^3$. The two 'transverse' polarizations $\epsilon_{\pm 1}$, for which the 3-vector $\boldsymbol{\epsilon}$ is perpendicular to the momentum $\boldsymbol{k}$, obey this condition, but of the $h = 0$ states, only the longitudinal combination

$$\epsilon_{\mathrm{L}} = \epsilon_0 + \epsilon_0' = \begin{pmatrix}1\\0\\0\\1\end{pmatrix} \tag{7.117}$$

does so. This longitudinal polarization vector is $\epsilon_{\mathrm{L}}^\mu = k^\mu/k$ and the corresponding plane wave can be written as $A_\mu(x) = \partial_\mu \theta(x)$, where $\theta(x) = (\mathrm{i}/k)\mathrm{e}^{-\mathrm{i}k\cdot x}$. It is thus a 'pure gauge', in the sense that it can be reduced to $A_\mu = 0$ by a gauge transformation, and is not physically meaningful. We see, then, that a photon exists only in the two polarization states with helicity $h = \pm 1$. In terms of classical light waves, this corresponds to the familiar fact that plane-wave solutions to Maxwell's equations have electric and magnetic fields transverse to the direction of propagation; the two helicity states correspond to states of right- and left-circular polarization in the classical theory.

Quantum field operators for spin-1 particles can be constructed from creation and annihilation operators for the allowed spin polarization states, but this is not entirely straightforward, owing to the fact that A_μ has more components than there are independent physical states. An immediate difficulty can be seen from exercise 3.6, which shows that the momentum conjugate to A_0 vanishes identically, which is inconsistent with commutation relations such as (7.14). Methods of circumventing this problem are described in many books on quantum field theory, but I do not propose to enter into them here. Instead, I shall discuss in chapter 9 an alternative approach to the quantization of gauge-invariant theories, namely the path-integral formalism, which is more convenient for many practical purposes.

7.6.2 Gravitons

The general-relativistic theory of gravity asserts, as we saw in chapter 4, that the metric tensor $g_{\mu\nu}(x)$ is not fixed, as in Minkowski spacetime, but is a dynamical quantity analogous in some respects to electromagnetic fields. We might therefore suspect the existence of gravitational radiation, which would be in some respects analogous to electromagnetic radiation. This is indeed a prediction of general relativity, and there exists a well-developed theory of the properties of gravitational waves and how they might be generated and detected. Here, I have the space only to deal with a few basic features, which bear directly on the possibility of finding a comprehensive quantum-mechanical theory of the physical world.

The equations which, in general relativity, serve a purpose analogous to that of Maxwell's equations in electromagnetism are the field equations (4.17), and our first objective is to derive from these a wave equation of the same kind as (7.109) or (7.110). To do this, we write the metric tensor as $g_{\mu\nu}(x) = g^{(B)}_{\mu\nu}(x) + h_{\mu\nu}(x)$, where $g^{(B)}_{\mu\nu}(x)$ is a 'background' metric describing the overall geometrical structure of whatever spacetime interests us, while $h_{\mu\nu}(x)$ is a small correction that we hope to interpret as a gravitational wave propagating through this background spacetime. By expanding the field equations to linear order in $h_{\mu\nu}(x)$, we obtain an approximate wave equation that describes freely-propagating waves. The essential features can be found most easily by taking the background to be Minkowski spacetime, so we take $g^{(B)}_{\mu\nu}(x) = \eta_{\mu\nu}$. Just as we obtained (7.109) by ignoring charged particles, so here we will ignore the presence of matter (which would, however, be necessary to generate the waves in the first place), setting $T_{\mu\nu} = 0$. As we found when obtaining the Schwarzschild solution, the field equations can then be written simply as $R_{\mu\nu} = 0$.

With $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, the affine connection coefficients $\Gamma^{\mu}_{\ \nu\sigma}$ can be approximated as in (4.6) and they are linear in $h_{\mu\nu}$. Therefore, our wave equation is derived only from the terms in the Ricci tensor (2.36) that are linear in Γ. That is

$$R_{\mu\nu} \approx \Gamma^{\lambda}_{\ \mu\nu,\lambda} - \Gamma^{\lambda}_{\ \mu\lambda,\nu} \approx 0. \tag{7.118}$$

Writing this out explicitly in terms of $h_{\mu\nu}$, we get

$$\Box h_{\mu\nu} + \partial_\mu \partial_\nu h^{\lambda}_{\lambda} - \partial^{\lambda}\left(\partial_\mu h_{\nu\lambda} + \partial_\nu h_{\mu\lambda}\right) = 0. \tag{7.119}$$

Since $h_{\mu\nu}$ is symmetric in μ and ν, it has ten independent components. At first sight, it might seem that a gravitational wave has ten possible polarizations, and thus that a quantum of gravitational energy, or *graviton*, would be a particle having ten independent spin polarizations, and obeying the rather complicated wave equation (7.119). At this point, however, we must take into account that the components of the Minkowski metric tensor have the special set of values $\eta_{\mu\nu}$ shown in (2.8) only when we use an inertial, Cartesian system of coordinates. There will be some sets of functions $h_{\mu\nu}(x)$ for which the metric $g_{\mu\nu} = \eta_{\mu\nu}+h_{\mu\nu}$

is exactly the metric of Minkowski spacetime, but expressed in a different coordinate system. An $h_{\mu\nu}$ of this kind does not describe a genuine gravitational wave. In this sense, general relativity has a gauge invariance symmetry quite similar to that of electromagnetism.

Let us work out the implications of this gauge invariance. Because we are considering only small changes $h_{\mu\nu}$ in the metric, we need consider only small changes in the coordinates, which can be specified by four small functions $\theta^\mu(x)$. That is, we consider the effect of changing to a new set of coordinates $\bar{x}^\mu = x^\mu + \theta^\mu(x)$. (Here I am using the same notation as in the derivation of (7.40), which makes sense only for small transformations.) According to the transformation laws that we obtained in §2.2, the components of the metric tensor in the new coordinates are

$$\bar{g}_{\mu\nu}(\bar{x}) = \frac{\partial x^\alpha}{\partial \bar{x}^\mu}\frac{\partial x^\beta}{\partial \bar{x}^\nu} g_{\alpha\beta}(x). \tag{7.120}$$

By expanding this to linear order in both $h_{\mu\nu}$ and θ^μ, we find that the new metric is $\bar{g}_{\mu\nu}(x) = \eta_{\mu\nu} + \bar{h}_{\mu\nu}(x)$, where

$$\bar{h}_{\mu\nu}(x) = h_{\mu\nu}(x) - \partial_\mu\theta_\nu(x) - \partial_\nu\theta_\mu(x). \tag{7.121}$$

Just as two vector potentials related by the gauge transformation (3.60) with an arbitrary function $\theta(x)$ represent the same physical situation, because they give the same electric and magnetic fields, so two small changes in the metric related by the gauge transformation (7.121) describe the same physical situation, because they give the same geometrical structure, albeit described in different coordinate systems.

Consider, in particular, the quantity $q_\mu = \partial^\lambda h_{\mu\lambda} - \frac{1}{2}\partial_\mu h^\lambda_\lambda$, whose derivatives appear in the wave equation (7.119). By making a gauge transformation, we find

$$\bar{q}_\mu(x) = q_\mu(x) - \Box\theta_\mu(x) \tag{7.122}$$

so we can arrange for $\bar{q}_\mu(x)$ to vanish, by choosing $\theta_\mu(x)$ to be solutions of the equation $\Box\theta_\mu = q_\mu$. Putting this another way, we are free to impose on the solutions of (7.119) the condition

$$\partial^\lambda h_{\mu\lambda} - \tfrac{1}{2}\partial_\mu h^\lambda_\lambda = 0 \tag{7.123}$$

which is the gravitational equivalent of the Lorentz gauge condition (7.111). It can be called the *harmonic gauge condition*, because it corresponds to choosing coordinates such that $g^{\mu\nu}\Gamma^\lambda_{\mu\nu} = 0$, which are called *harmonic coordinates*. In that case, the wave equation becomes simply $\Box h_{\mu\nu} = 0$, which is the Klein–Gordon equation for a massless particle.

We have now established that a graviton is a massless particle, which therefore travels with the speed of light. What about its spin? A plane-wave solution to (7.119) can be written as

$$h_k^{\mu\nu}(x) = \epsilon^{\mu\nu}\mathrm{e}^{-\mathrm{i}k\cdot x} \tag{7.124}$$

where $k_\mu k^\mu = 0$ and $\epsilon^{\mu\nu}$ is a symmetric polarization tensor with ten independent components. The gauge condition (7.123) implies that this tensor must obey the four equations

$$k_\lambda \epsilon^{\mu\lambda} = \tfrac{1}{2} k^\mu \epsilon^\lambda_\lambda \tag{7.125}$$

for $\mu = 0, \ldots, 3$ and this reduces the number of independent components to six. We can, however, make a further gauge transformation, using

$$\theta^\mu(x) = \theta^\mu \mathrm{e}^{-\mathrm{i}k\cdot x} \tag{7.126}$$

where θ^μ are four arbitrary constants. The new solution (which is physically equivalent to the old one) is $\bar{h}^{\mu\nu}(x) = \bar{\epsilon}^{\mu\nu}\mathrm{e}^{-\mathrm{i}k\cdot x}$, where

$$\bar{\epsilon}^{\mu\nu} = \epsilon^{\mu\nu} - k^\mu\theta^\nu - k^\nu\theta^\mu. \tag{7.127}$$

Using the fact that $k_\mu k^\mu = 0$, it is easy to check that $\bar{h}^{\mu\nu}(x)$ still obeys the harmonic gauge condition. Because of the four arbitrary constants, the number of physically meaningful independent components in $\bar{\epsilon}^{\mu\nu}$ is now $6-4=2$. Because the graviton is massless, these must correspond to two states of opposite helicity. The values of this helicity can be determined by the same methods we have used in previous cases. The algebra is a little more complicated, though, so I shall just quote the result, which is that $h = \pm 2$. Thus, the graviton is a massless spin-2 particle. As for the photon, however, some of the helicity states that we might have expected (namely $h = 0, \pm 1$) correspond to gauge degrees of freedom and not to genuine particle states.

While photons and massive spin-1 particles are routinely detected by experimenters, no graviton has yet been observed. Quite possibly, this is because no sufficiently sensitive detector has yet been constructed. There is, on the other hand, some rather compelling, though indirect, evidence for the existence of classical gravitational waves. This comes from observations of a single astronomical object—a binary pulsar discovered by R. A. Hulse and J. H. Taylor in the 1970s. The orbital frequency of this binary star system has been found to be increasing, in a manner that can be attributed to energy loss through the emission of gravitational radiation. Indeed, the frequency has been accurately monitored over many years and is found to agree remarkably well with theoretical predictions based on this interpretation.

7.7 Wave Equations in Curved Spacetime

It ought, of course, to be possible to study wave equations and field theories in curved spacetimes. When this is done, it turns out that there are difficulties of interpretation over and above those we have already encountered in Minkowski spacetime, and these difficulties have not, to my mind, been completely resolved. More detailed discussions than I can give here can be found, for example, in Birrell and Davies (1982) and Wald (1984).

The first requirement, obviously, is that wave equations should be covariant, and therefore the action should be invariant, under general coordinate transformations. Starting from the theories we have already considered, two steps are necessary to construct suitable actions: we must use the covariant spacetime volume element (4.12) and replace partial derivatives with covariant derivatives. It is also possible to add further terms involving the Riemann curvature tensor, which will vanish if the spacetime happens to be flat. In the case of a scalar field, these steps can be carried out straightforwardly. The covariant derivative of a scalar quantity is the same as the partial derivative, and we arrive at an action of the form

$$S = \int \mathrm{d}^4x\, (-g(x))^{1/2} \left[g^{\mu\nu}(x)\partial_\mu\phi^*\partial_\nu\phi - m^2\phi^*\phi + \xi R(x)\phi^*\phi \right] \tag{7.128}$$

where $R(x)$ is the Ricci curvature scalar defined in (2.51) and ξ is a dimensionless number. This additional term is the only possible one that does not involve dimensionful coefficients. The corresponding Euler–Lagrange equation is

$$g^{\mu\nu}\nabla_\mu\nabla_\nu\phi + \left(m^2 - \xi R\right)\phi = 0 \tag{7.129}$$

where ∇_μ is the covariant derivative. (Recall that, although $\nabla_\mu\phi = \partial_\mu\phi$, this quantity is a vector, which must be acted on with a covariant derivative.) To derive (7.129), we use an integration by parts, and the covariant derivative enters through the covariant version of Gauss' theorem exhibited as equation (A.23) of appendix A. The value of ξ is not determined by any known physical principle and, since the effects of spacetime curvature are too small to measure in the laboratory, it cannot be determined by experiment either. The case $\xi = 0$ is called *minimal coupling*, for obvious reasons. An interesting case is $\xi = \frac{1}{6}$. If $\xi = \frac{1}{6}$ and $m = 0$, the theory possesses a symmetry known as *conformal invariance*. A conformal transformation means replacing the metric $g_{\mu\nu}(x)$ with $\Omega(x)^2 g_{\mu\nu}(x)$, where $\Omega(x)$ is an arbitrary function. If at the same time we replace $\phi(x)$ with $\Omega^{-1}(x)\phi(x)$, then it can be shown that the wave equation (7.129) is unchanged. Whether we should expect this symmetry to be respected by nature is not clear. At any rate, the case $\xi = \frac{1}{6}$ is known as *conformal coupling*.

To construct a generally covariant version of the Dirac equation requires rather more thought. We have seen that spinor wavefunctions do not have the same transformation properties as any of the tensors considered in chapter 2, so we do not yet know how to form their covariant derivatives. In order to do this, we first recall that it is always possible to set up a system of locally inertial Cartesian coordinates, valid in a sufficiently small region of spacetime. Strictly speaking, this must be an infinitesimal region surrounding, say, the point X with coordinates $x^\mu = X^\mu$. I shall denote these local coordinates by y^a, using Latin indices $a, b, c, \ldots$ to distinguish them from the large-scale coordinates x^μ. In terms of these coordinates, the metric tensor is given at X by the Minkowski form η_{ab}. I shall denote the transformation matrix Λ (equation(2.13)) which relates the two sets of coordinates by the special symbol e:

$$e^{\mu}_{\ a}(X) = \left.\frac{\partial x^{\mu}}{\partial y^{a}}\right|_{x^{\mu}=X^{\mu}} \qquad \text{and} \qquad e^{a}_{\ \mu}(X) = \left.\frac{\partial y^{a}}{\partial x^{\mu}}\right|_{x^{\mu}=X^{\mu}} . \tag{7.130}$$

If we set up a locally inertial frame of reference at each point of spacetime, in such a way that the directions of their axes vary smoothly from one point to another, then we obtain a set of four vector fields $e^{\mu}_{\ 0}(x), \ldots, e^{\mu}_{\ 3}(x)$ which specify, at each point, the directions of these axes. This set of vector fields is known variously as a *vierbein* (a German expression meaning 'four legs'), a *tetrad*, or a *frame field*. In theories that envisage numbers of spacetime dimensions other than four, it is called a *vielbein* (a German expression meaning 'many legs'). At a given point X, the vierbein constitutes a set for four 4-vectors $e^{a}_{\ 0}(X), \ldots, e^{a}_{\ 3}(X)$, which specify the directions and scales of the large-scale coordinates relative to the inertial coordinates at X. Considered as a whole, the 16 components of the vierbein constitute a kind of rank-2 tensor field whose μ indices transform as a vector under general coordinate transformations and can be raised and lowered using g, while its a indices transform as a 4-vector under Lorentz transformations in the local coordinates and can be raised and lowered using η. By construction, the vierbein satisfies the relations

$$e^{\mu}_{\ a}(x)e^{a}_{\ \nu}(x) = \delta^{\mu}_{\nu} \qquad e^{a}_{\ \mu}(x)e^{\mu}_{\ b}(x) = \delta^{a}_{b} \tag{7.131}$$

and

$$e^{\mu}_{\ a}(x)e^{\nu}_{\ b}(x)g_{\mu\nu}(x) = \eta_{ab} \qquad e^{a}_{\ \mu}(x)e^{b}_{\ \nu}(x)\eta_{ab} = g_{\mu\nu}(x). \tag{7.132}$$

Its 16 independent components evidently carry two kinds of information. First, as we see from (7.132), they contain all the information needed to construct the 10 independent components of the metric tensor field. Second, each local inertial frame can be redefined by Lorentz transformations, involving three independent rotations and three boosts, and the remaining six degrees of freedom in the vierbein specify the choices we have actually made.

It is now possible to describe any vector quantity either in terms of its components $V^{\mu}(x)$ relative to the large-scale coordinate directions, which I shall refer to for brevity as a *coordinate vector*, or in terms of its components $V^{a}(x)$ relative to the local coordinate directions at x, which I shall call a *Lorentz vector*. The two sets of components are obviously related by

$$V^{\mu}(x) = e^{\mu}_{\ a}(x)V^{a}(x) \qquad \text{and} \qquad V^{a}(x) = e^{a}_{\ \mu}(x)V^{\mu}(x). \tag{7.133}$$

In fact, any tensor field can be expressed in terms of components with any combination of a-type and μ-type indices we happen to find convenient. The advantage of this is clear: we know how to deal with spinors in the local inertial coordinates, and the vierbein permits us to embed these in the curved spacetime. In order to work out the covariant derivative of a spinor, we need a suitable rule for parallel transport. We shall first work out the rule for a Lorentz vector, which will apply, for example, to the current $\bar{\psi}(x)\gamma^a\psi(x)$, and then deduce the corresponding rule for the spinor itself.

To transport $V^a(x)$ to the point $x + \mathrm{d}x$, we need only to translate (2.23) into the language of locally inertial coordinates. The transported vector will be given by

$$V^a(x \to x + \mathrm{d}x) = V^a(x) - \omega^a_{\ b\nu}(x)V^b(x)\mathrm{d}x^\nu \tag{7.134}$$

where the coefficients $\omega^a_{\ b\nu}(x)$ are the components of what is called the *spin connection*. They involve both the affine connection, which defines parallel transport of the vector itself, and the vierbein, which relates the locally inertial coordinates at x to those at $x + \mathrm{d}x$. We use the relations

$$V^\mu(x) = e^\mu_{\ a}(x)V^a(x) \qquad V^\mu(x \to x + \mathrm{d}x) = e^\mu_{\ a}(x + \mathrm{d}x)V^a(x \to x + \mathrm{d}x)$$

together with the expansion $e^\mu_{\ a}(x+\mathrm{d}x) \simeq e^\mu_{\ a}(x)+e^\mu_{\ a,\nu}(x)\mathrm{d}x^\nu$ to convert (7.134) into a transport equation for V^μ and compare the result with (2.23). We find that the spin connection is given by

$$\omega^a_{\ b\nu} = e^a_{\ \mu}e^\mu_{\ b,\nu} + e^a_{\ \mu}e^\sigma_{\ b}\Gamma^\mu_{\ \sigma\nu}. \tag{7.135}$$

With the spin connection in hand, we can generalize (2.28) to obtain the covariant derivative of a tensor field with both a- and μ-type indices, including a Γ term for each coordinate index and an ω term for each Lorentz index. The vierbein itself is such a tensor, and by rearranging (7.135) we see that its covariant derivative vanishes:

$$\nabla_\nu e^\mu_{\ a} = e^\mu_{\ a,\nu} + \Gamma^\mu_{\ \sigma\nu}e^\sigma_{\ a} - \omega^b_{\ a\nu}e^\mu_{\ b} = 0. \tag{7.136}$$

This result should give alert readers pause for thought. We saw in §2.3.5 that, in order to make the notion of parallel transport as defined by the affine connection consistent with that defined by the metric, the covariant derivative of the metric should vanish, and it does so only when the affine connection is the metric connection (2.50). Although we shall usually want Γ to be this metric connection, we have not actually assumed this in order to derive (7.136). To resolve this point to their own satisfaction, readers may like to consider the conditions under which any two of the notions of parallelism defined by the affine connection, the metric and the vierbein become equivalent. In particular, consideration of the covariant derivatives of $g_{\mu\nu}$, of η_{ab} and of equations (7.132) should prove illuminating. Let us impose the consistency condition that the magnitude of a transported Lorentz vector should be preserved, so that

$$\eta_{ab}V^a(x \to x + \mathrm{d}x)V^b(x \to x + \mathrm{d}x) = \eta_{ab}V^a(x)V^b(x).$$

It is easy to see that the spin connection must be antisymmetric, in the sense that

$$\omega_{ab\nu}(x) \equiv \eta_{ac}\omega^c_{\ b\nu}(x) = -\omega_{ba\nu}(x). \tag{7.137}$$

By using this condition, readers should be able to show that (2.48) is satisfied, so Γ must be the metric connection.

We can now turn our attention to spinors, which should satisfy a rule for parallel transport of the form

$$\psi(x \to x + \mathrm{d}x) = \psi(x) - \Omega_\nu(x)\psi(x)\mathrm{d}x^\nu \tag{7.138}$$

where $\Omega_\nu(x)$ is a suitable connection coefficient. This coefficient, like the previous ones, has three indices; the first two are those it possesses by virtue of being a 4 × 4 spin matrix. To discover what this coefficient is, we demand that, in particular, the scalar quantity $S(x) = \bar{\psi}(x)\psi(x)$ should be invariant under parallel transport, while the Lorentz vector $V^a(x) = \bar{\psi}(x)\gamma^a\psi(x)$ should be transported according to (7.134). From (7.138), we find

$$S(x \to x + \mathrm{d}x) = S(x) - \bar{\psi}(x)\left[\gamma^0\Omega_\nu^\dagger(x)\gamma^0 + \Omega_\nu(x)\right]\psi(x)\mathrm{d}x^\nu \tag{7.139}$$

so our first condition gives

$$\gamma^0\Omega_\nu^\dagger(x)\gamma^0 = -\Omega_\nu(x). \tag{7.140}$$

Using this, we find similarly that $V^a(x)$ is correctly transported provided that

$$[\gamma^a, \Omega_\nu(x)] = \omega^a{}_{b\nu}(x)\gamma^b. \tag{7.141}$$

Taking into account the antisymmetry property (7.137), we can use (7.35) and (7.51) to identify the matrix satisfying these two conditions as

$$\Omega_\nu(x) = -\tfrac{\mathrm{i}}{4}\omega_{ab\nu}(x)\sigma^{ab} = \tfrac{1}{8}\omega_{ab\nu}(x)[\gamma^a, \gamma^b]. \tag{7.142}$$

Then the covariant derivative of the spinor is

$$\nabla_\nu\psi(x) = [\partial_\nu + \Omega_\nu(x)]\,\psi(x). \tag{7.143}$$

It is now a straightforward matter to write down the covariant version of the Dirac equation (7.23). The γ matrices are valid only within the local inertial frame and must be contracted with the covariant derivative by using the vierbein:

$$\left[\mathrm{i}e^\mu_a(x)\gamma^a\nabla_\mu - m\right]\psi(x) = 0. \tag{7.144}$$

We can tidy this up by defining a set of covariant γ matrices

$$\gamma^\mu(x) = e^\mu_a(x)\gamma^a \tag{7.145}$$

and it may easily be verified that these satisfy the generally covariant version of the Clifford algebra condition (7.26):

$$\{\gamma^\mu(x), \gamma^\nu(x)\} = 2g^{\mu\nu}(x). \tag{7.146}$$

The generally covariant action is clearly

$$S = \int \mathrm{d}^4x\, (-g(x))^{1/2}\bar{\psi}(x)\left[\mathrm{i}\gamma^{\mu}(x)\nabla_{\mu} - m\right]\psi(x) \tag{7.147}$$

and in this case no curvature term can be added with a dimensionless coefficient. If we wish, we can express $(-g)^{1/2}$ as $\det(e^{a}_{\ \mu})$.

Clearly, wave equations such as (7.129) and (7.144) do not, in general, have simple plane-wave solutions. Only in very special cases, indeed, can their solutions be found in closed form. When we try to reinterpret these equations in terms of quantum fields, we encounter a new difficulty of principle. As a matter of fact, this difficulty really exists for quantum field theories in Minkowski spacetime as well, though we do not need to worry about it for most practical purposes. A simply-stated fact that illustrates the Minkowski version of the problem is this: *the vacuum state, which from the point of view of an inertial observer contains no particles, will be perceived by an accelerating observer as containing a thermal bath of particles at a temperature proportional to the observer's acceleration*. Proving this remarkable fact is not quite so simple, but I shall outline one of several standard calculations that illustrate its truth. We consider the theory of massless spin-0 particles in a Minkowskian spacetime that has only two dimensions. In an inertial frame of reference, we can use coordinates x and t. Another set of coordinates, ξ and η, invented by W Rindler, is related to these by

$$x = \alpha^{-1}\mathrm{e}^{\alpha\xi}\cosh(\alpha\eta) \qquad t = \alpha^{-1}\mathrm{e}^{\alpha\xi}\sinh(\alpha\eta) \tag{7.148}$$

where α is a constant. We should note at once that, although both ξ and η are allowed to vary between $-\infty$ and $+\infty$, these coordinate values cover only part of the whole spacetime, namely the region $x > |t|$, which is called the *Rindler wedge*. In this region, the line element is given by

$$\mathrm{d}\tau^2 = \mathrm{d}t^2 - \mathrm{d}x^2 = \mathrm{e}^{2\alpha\xi}\left(\mathrm{d}\eta^2 - \mathrm{d}\xi^2\right). \tag{7.149}$$

To see the meaning of these coordinates, consider an observer whose ξ coordinate is fixed. Relative to the inertial frame of reference, the equation of his path through spacetime is $x^2 - t^2 = a_{\mathrm{p}}^2$, where $a_{\mathrm{p}} = \alpha\exp(-\alpha\xi)$. A little algebra suffices to show that his velocity $u = \mathrm{d}x/\mathrm{d}t$ and his acceleration $a = \mathrm{d}^2x/\mathrm{d}t^2$ obey the relation

$$a_{\mathrm{p}} = \left(1 - u^2\right)^{-3/2} a. \tag{7.150}$$

Compare this with the result of exercise 2.2, setting $c = 1$ and $v = u$. We see that a_{p} is the observer's *proper acceleration*; that is, his acceleration relative to an inertial frame of reference in which he is instantaneously at rest. For this observer, proper time is given by (7.149) with $\mathrm{d}\xi = 0$ as $\mathrm{d}\tau = \exp(\alpha\xi)\mathrm{d}\eta$.

Suppose that $\hat{\phi}(x,t)$ is an Hermitian scalar field, describing particles that are their own antiparticles. According to (7.11), with only one space dimension and

$\omega(q) = |q|$ for massless particles of momentum q, it can be expressed in terms of creation and annihilation operators as

$$\hat{\phi}(x,t) = \int \frac{\mathrm{d}q}{(2\pi)2|q|} \left[\hat{a}(q)\mathrm{e}^{-\mathrm{i}|q|t+\mathrm{i}qx} + \hat{a}^\dagger(q)\mathrm{e}^{\mathrm{i}|q|t-\mathrm{i}qx}\right]. \tag{7.151}$$

Now, as long as we consider the field inside the Rindler wedge, the Klein–Gordon equation is

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}\right)\phi = \mathrm{e}^{-2\alpha\xi}\left(\frac{\partial^2}{\partial \eta^2} - \frac{\partial^2}{\partial \xi^2}\right)\phi = 0. \tag{7.152}$$

In terms of ξ and η, this equation has plane wave solutions, though they are by no means the same as the ones that appear in (7.151). For example

$$\phi_k(\xi,\eta) = \mathrm{e}^{-\mathrm{i}k(\eta-\xi)} = [\alpha(x-t)]^{\mathrm{i}k/\alpha}. \tag{7.153}$$

Therefore, the field can also be expanded as

$$\hat{\phi}_\mathrm{R}(\xi,\eta) = \int \frac{\mathrm{d}k}{(2\pi)2|k|} \left[\hat{b}(k)\mathrm{e}^{-\mathrm{i}|k|\eta+\mathrm{i}k\xi} + \hat{b}^\dagger(k)\mathrm{e}^{\mathrm{i}|k|\eta-\mathrm{i}k\xi}\right] \tag{7.154}$$

where the subscript R reminds us that this is valid only inside the Rindler wedge. The new creation and annihilation operators $\hat{b}^\dagger(k)$ and $\hat{b}(k)$ obey the same commutation relation as $\hat{a}^\dagger(q)$ and $\hat{a}(q)$, but they are not the same operators. In fact, we can use (7.12) to write

$$\hat{b}(k) = \int_{-\infty}^{\infty} \mathrm{d}\xi\ \mathrm{e}^{\mathrm{i}|k|\eta-\mathrm{i}k\xi}\left(|k|\phi + \mathrm{i}\partial\phi/\partial\eta\right) \tag{7.155}$$

and use the expression (7.151) for ϕ to find $\hat{b}(k)$ in terms of $\hat{a}(q)$ and $\hat{a}^\dagger(q)$. (However, because $\hat{\phi}(x,t)$ exists throughout the Minkowski spacetime while $\hat{\phi}_\mathrm{R}(\xi,\eta)$ exists only in the Rindler wedge, it is not possible to express $\hat{a}(q)$ in terms of $\hat{b}(k)$ and $\hat{b}^\dagger(k)$ alone.) Let us write the result as

$$\hat{b}(k) = \int \mathrm{d}q \left[\alpha_k(q)\hat{a}(q) + \beta_k(q)\hat{a}^\dagger(q)\right] \tag{7.156}$$

where the functions $\alpha_k(q)$ and $\beta_k(q)$ are defined by integrals that are somewhat awkward to compute. A relation of this kind between the creation and annihilation operators associated with different sets of solutions to a wave equation is called a *Bogoliubov transformation*. It is interesting to calculate the expectation value $N(k,k') = \langle 0|\hat{b}^\dagger(k)\hat{b}(k')|0\rangle$, where $|0\rangle$ is the Minkowski-spacetime vacuum state, for which

$$\hat{a}(q)|0\rangle = 0 \qquad \text{and} \qquad \langle 0|\hat{a}^\dagger(q) = 0. \tag{7.157}$$

According to (7.22) the quantity

$$\frac{N(k,k)\mathrm{d}k}{(2\pi)2|k|} \tag{7.158}$$

gives the number of particles with momentum between k and $k + \mathrm{d}k$ as seen by a Rindler observer. However, if there is a finite number of particles per unit volume in an infinite volume, then $N(k, k)$ will be infinite. By first calculating $N(k, k')$, we will be able to extract a finite answer for the number of particles per unit volume. It is given by

$$\begin{aligned} N(k, k') &= \int \mathrm{d}q\mathrm{d}q'\, \beta_k^*(q)\beta_{k'}(q')\langle 0|\hat{a}(q)\hat{a}^\dagger(q')|0\rangle \\ &= \int \mathrm{d}q\mathrm{d}q'\, \beta_k^*(q)\beta_{k'}(q')(2\pi)2|q|\delta(q-q') \\ &= 4\pi \int \mathrm{d}q\, |q|\beta_k^*(q)\beta_{k'}(q). \end{aligned}$$

In the first line of this calculation, three of the terms that result from calculating $\hat{b}^\dagger(k)\hat{b}(k')$ have disappeared because of the conditions (7.157) that define the vacuum state. In the second line, I have used the commutator (7.17)—adjusted to one space dimension—and (7.157) again. The remaining integral can be evaluated (though this is not entirely straightforward) with the result

$$N(k, k') = \langle 0|\hat{b}^\dagger(k)\hat{b}(k')|0\rangle = (2\pi)2|k|\delta(k-k')\left(\mathrm{e}^{2\pi|k|/\alpha} - 1\right)^{-1}. \quad (7.159)$$

On setting $k' = k$, the infinite factor $\delta(0)$ can be interpreted, as discussed in appendix D, as representing the infinite volume. The factor $\left(\mathrm{e}^{2\pi|k|/\alpha} - 1\right)^{-1}$ has the same form as the Bose–Einstein occupation number, to be discussed in chapter 10 (see equation (10.64)), for a gas of bosonic particles. In particular, the argument of the exponential, $2\pi|k|/\alpha$, corresponds in thermodynamic language to $\epsilon(k)/k_\mathrm{B}T$, where $\epsilon(k)$ is the energy of a particle of momentum k, T is the temperature of the gas and k_B is Boltzmann's constant.

To interpret this result correctly, recall from our earlier discussion that for an observer whose ξ coordinate is fixed, say at $\xi = \xi_0$, proper time is measured by $\tau = \mathrm{e}^{\alpha\xi_0}\eta$. In terms of this proper time, we can write a positive-energy plane wave as

$$\exp(-\mathrm{i}|k|\eta + \mathrm{i}k\xi) = \exp\left(-\mathrm{i}|k|\mathrm{e}^{-\alpha\xi_0}\tau + \mathrm{i}k\xi\right) \quad (7.160)$$

which will be interpreted by this observer as corresponding to a particle of energy $\epsilon(k) = \mathrm{e}^{-\alpha\xi_0}|k|$. From this observer's point of view, then, we must identify $2\pi|k|/\alpha = \epsilon(k)/k_\mathrm{B}T$, where $k_\mathrm{B}T = a_\mathrm{p}/2\pi$ and $a_\mathrm{p} = \alpha\mathrm{e}^{-\alpha\xi_0}$ is the observer's proper acceleration. From the point of view of statistical mechanics, this is at first sight a strange result. As we shall see in more detail in chapter 10, the Bose–Einstein distribution normally arises as an *ensemble average* of the number of particles occupying a given quantum state; that is, an average over all the microscopic states that are compatible with given values of a small number of macroscopic quantities such as temperature. In the process of taking this average, a great deal of information about the detailed state is lost. By contrast, the expectation value in (7.159) is taken in a pure quantum state, which contains all the information that quantum mechanics allows us to have. One way of understanding this is to remember that the field ϕ_{R}, which contains all the information available to an accelerated Rindler observer exists only in the Rindler wedge. The lost information concerns the state of the field in the rest of the spacetime, to which the observer has no access.

From the point of view of quantum field theory, the result is strange for a different reason. It tells us that the number of particles present in a given quantum state depends on the frame of reference from which the state is observed. Indeed, the very concept of a 'particle' has turned out to have no frame-independent meaning. Nor, therefore, does the notion of a 'vacuum' as the state in which no particles are present. We could perfectly well define a new vacuum state, say $|\bar{0}\rangle$, in Minkowski spacetime by requiring that $\hat{b}(k)|\bar{0}\rangle = 0$ instead of (7.157). From the point of view of an inertial observer, this would be a state in which particles were present. So long as we deal only with Minkowski spacetime, this ambiguity can be ignored for most practical purposes. The number of particles present in a given state is the same when the state is observed from any *inertial* frame and we normally take inertial frames of reference to define a preferred concept of 'particle' and a preferred vacuum state. To be sure, our practical frames of reference, such as those fixed on the earth's surface, are not exactly inertial. To estimate the likely effect of this, let us calculate the temperature corresponding to an acceleration g, that due to gravity at the earth's surface. To get an answer in laboratory units, we must use dimensional analysis to reinstate the appropriate factors of $\hbar$ and c. The result is $T = g\hbar/2\pi k_{\mathrm{B}}c \simeq 10^{-20}\,\mathrm{K}$, so the effect is completely negligible.

When we deal with a curved spacetime, this option is no longer open to us, because no one set of inertial coordinates can, in general, cover the whole spacetime. In particular, there will in general exist no quantum state that will appear to every inertial observer to be devoid of particles. In view of the equivalence principle, indeed, we might expect to be able to recast the general idea that an accelerating observer in a vacuum observes the presence of particles as a statement to the effect that particles can be created by a gravitational field. A celebrated result of Hawking (1974) confirms this in the case of a black hole. According to Hawking's analysis, an observer far from a black hole, whose spatial coordinates (r, θ, ϕ) (in the notation of §4.4 and §4.5) are fixed, will observe particles traveling outwards. This is true, at least, if we specify the quantum state by assuming that there is no radiation coming in from infinity. For a black hole of mass M, this *Hawking radiation* is the same as that emitted by a black body whose temperature in natural units is $T = g_M/2\pi k_{\mathrm{B}}$. The acceleration g_M here is

$$g_M = \frac{1}{4GM} = \frac{GM}{r_{\mathrm{S}}^2} \tag{7.161}$$

where r_{S} is the Schwarzschild radius. In Newtonian terms, this is the acceleration due to gravity at the surface of a body of mass M and radius r_{S}, so we might loosely identify it as the acceleration due to gravity at the event horizon, although the proper acceleration of an object fixed at $r = r_S$ is infinite. It is still true in the quantum theory that particles cannot escape from inside the event horizon, so the radiated particles must be thought of as being created by the gravitational field outside the event horizon. Nevertheless, the energy to create them must be provided by a reduction in the mass of the black hole. In fact, a detailed analysis shows that the energy density near the event horizon is *negative* (a quantum effect that has no classical counterpart) and that the hole's mass decreases because of an inward flow of negative energy. In this way, it seems that a black hole can 'evaporate' and perhaps might eventually disappear altogether. However, the theories we have thought about in this section concern quantum field theory in a spacetime with a predetermined geometry. To determine the fate of a black hole, we need to know about the 'backreaction' of the quantum field on the black hole geometry. To the best of my knowledge, no entirely reliable way of doing this has yet been found. A reasonable hypothesis seems to be that the energy $E = Mc^2$ should decrease roughly in accordance with Stefan's law $\mathrm{d}(Mc^2)/\mathrm{d}t = -\sigma A T^4$, where A is the surface area of a black body and σ is the Stefan–Boltzmann constant (see equation (10.90)), perhaps modified to take account of the radiation of particles other than photons. Taking A to be the area of the event horizon, we find that $\mathrm{d}M/\mathrm{d}t = -\text{constant} \times M^{-2}$. Thus, the rate of evaporation is very small for a large black hole, but becomes explosive as a small black hole nears the end of its life. It is not hard to work out that the life expectancy of, say, a black hole of one solar mass is of the order of 10^{70} years—vastly longer than the present age of the universe, which is around 10^{10} years. It has been speculated that small black holes, created in the very early universe, might be exploding around now, contributing to the observed flux of cosmic rays, but it is hard to test this idea in any stringent way.

Exercises

7.1. In the Lagrangian density (7.7), let $\phi = 2^{-1/2}(\phi_1 + \mathrm{i}\phi_2)$, where ϕ_1 and ϕ_2 are real, and show that $\mathcal{L}$ becomes the sum of independent terms for ϕ_1 and ϕ_2. Identify the two conjugate momenta and carry out the canonical quantization procedure. Show that ϕ_1 and ϕ_2 are the field operators for two particle species, each of which is its own antiparticle. Verify that your commutation relations agree with (7.14) and (7.15) when ϕ is expressed in terms of ϕ_1 and ϕ_2. How are the type 1 and type 2 particle states related to the particle and antiparticle states of §7.2? How does the factor of $2^{-1/2}$ affect the definition of the conjugate momenta, the commutation relations, the definition of creation and annihilation operators and the normalization of particle states?

7.2. Let γ^μ be a set of matrices satisfying (7.26), (7.48) and (7.50) and let U be any constant unitary matrix. Show that the four matrices $U\gamma^\mu U^{-1}$ also have these properties and can therefore be used in the Dirac equation.

7.3. For any 4-vector a^μ, show that $\not{a}\not{a} = a_\mu a^\mu$.

7.4. The spinors (7.68) and (7.74) give plane-wave solutions of the Dirac equation in the rest frame, when the γ matrices (7.27) are used. Denote them by $u(m, s)$ and $v(m, s)$. Show that, in a frame where the momentum is k^μ, the spinors $u(k, s) = (k^0+m)^{-1/2}(\not{k}+m)u(m, s)$ and $v(k, s) = (k^0+m)^{-1/2}(-\not{k}+m)v(m, s)$ give plane-wave solutions which satisfy the orthonormality conditions (7.81) and (7.82).

Use the relations (7.83) and (7.84) to verify that the anticommutation relations (7.86) for creation and annihilation operators follow from the anticommutator (7.87) of the field and its conjugate momentum (7.78).

7.5. The idea of charge conjugation requires that $(\psi^{\mathrm{c}})^{\mathrm{c}} = \eta\psi$, where η is a constant phase factor ($|\eta| = 1$). Why is this? Assuming that $\eta = 1$, show that $\mathcal{C}\mathcal{C}^* = 1$ and $CC^* = -1$ where $\mathcal{C}$ and C are the charge conjugation matrices defined in §7.3.6. Do not assume that the γ matrices are those given in (7.27).

7.6. Show that $\gamma_\mu\gamma^\mu = 4$. Show that $[\gamma_\mu, \gamma_\tau]\gamma^5$ is proportional to $[\gamma_\nu, \gamma_\sigma]$, where (μ, ν, σ, τ) is some permutation of (0, 1, 2, 3).

Hence show that $[\gamma_\mu, \gamma_\tau]\gamma^5 = -\mathrm{i}\epsilon_{\mu\nu\sigma\tau}\gamma^\nu\gamma^\sigma$ and that the Pauli–Lubanski vector (7.44) can be expressed in the form (7.57).

7.7. If $S(\Lambda)$ is a Lorentz transformation matrix that satisfies (7.32), show that $S^{-1}(\Lambda)\gamma^5 S(\Lambda) = \det(\Lambda)\gamma^5$. (It may be helpful to read about the Levi-Civita symbol in appendix A.)

7.8. If the *chiral projection operators* are defined as $P_{\mathrm{R}} = \frac{1}{2}(1 + \gamma^5)$ and $P_{\mathrm{L}} = \frac{1}{2}(1 - \gamma^5)$, show that $P_{\mathrm{R}}^2 = P_{\mathrm{R}}$, $P_{\mathrm{L}}^2 = P_{\mathrm{L}}$ and $P_{\mathrm{R}}P_{\mathrm{L}} = P_{\mathrm{L}}P_{\mathrm{R}} = 0$. If $\psi_{\mathrm{L}} = P_{\mathrm{L}}\psi$, show that $\bar{\psi}_{\mathrm{L}} = \bar{\psi}P_{\mathrm{R}}$. Show that the charge conjugate of a left-handed spinor is right handed and *vice versa*.

7.9. If $\psi = \psi_{\mathrm{L}} + \psi_{\mathrm{R}}$, show that $\bar{\psi}\psi = \bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}} + \bar{\psi}_{\mathrm{R}}\psi_{\mathrm{L}}$ and that $\bar{\psi}\not{\partial}\psi = \bar{\psi}_{\mathrm{L}}\not{\partial}\psi_{\mathrm{L}} + \bar{\psi}_{\mathrm{R}}\not{\partial}\psi_{\mathrm{R}}$.

7.10. In the standard representation of the γ matrices (7.27) show that the transpose of the charge conjugation matrix C is $C^{\mathrm{T}} = -C$. Now define the charge conjugate of the vector current $V^\mu = \bar{\psi}\gamma^\mu\psi$ to be $V^{\mathrm{c}\mu} = \bar{\psi}^{\mathrm{c}}\gamma^\mu\psi^{\mathrm{c}}$. Show that $V^{\mathrm{c}\mu} = +V^\mu$ if the components of ψ are treated as ordinary numbers and $V^{\mathrm{c}\mu} = -V^\mu$ if they are regarded as anticommuting Grassmann numbers. Which treatment is more appropriate in view of the antiparticle interpretation?

Chapter 8

Forces, Connections and Gauge Fields

One of the central problems faced by theoretical physics is to explain the nature and origin of the forces that act between fundamental particles. In the case of gravity, this is elegantly achieved (at the non-quantum-mechanical level) by general relativity. With hindsight, we may say that an explanation of gravitational forces arises naturally—indeed, almost inevitably—from a systematic and explicit account of the geometrical structure of spacetime. The origin of gravitational forces, as described in chapter 4, may be summarized as follows:

(i) To relate physical quantities (represented by tensors) at different points of spacetime, we must introduce a specific geometrical structure, the affine connection, which defines parallel transport.

(ii) The simplest situation is that the connection coefficients are zero everywhere (or can be made so by a suitable choice of coordinates). Departures from this situation are what we recognize as gravitational forces.

(iii) It appears that the particular kinds of departure countenanced by nature can be embodied in a principle of least action.

In essence, *gravitational forces arise from communication between different points of spacetime*. At least at the level of description that accounts for all current experimental observations, it appears that all known forces can be considered to arise in essentially this way. In what follows, I shall first describe how this comes about in the case of electromagnetism, and then discuss how the idea can be generalized to encompass forces of other kinds.

8.1 Electromagnetism

Consider a particle described by a complex wavefunction

$$\phi(x) = \phi_1(x) + \mathrm{i}\phi_2(x). \tag{8.1}$$

The absolute phase of ϕ is not an observable quantity. If, for example, each wavefunction in (5.52) is multiplied by the same constant phase factor $\exp(\mathrm{i}\theta)$, this factor cancels out in the final result. On the other hand, variations of the phase through spacetime do have a physical significance, because a varying phase angle $\theta(x)$ is differentiated by the momentum operator. This may be expressed differently, if we think of the value of ϕ at the spacetime point x as a point in an 'internal space', namely the (ϕ_1, ϕ_2) plane. The fact that the phase of ϕ is unobservable implies that no particular direction in this plane has any special physical significance. To represent the whole function $\phi(x)$, we must erect a (ϕ_1, ϕ_2) plane at each point of spacetime. The geometrical structure which results is a *fibre bundle*. It is analogous to the Galilean spacetime fibre bundle, in which a three-dimensional Euclidean space is erected at each point in time, or to the tangent and cotangent bundles that we discussed in §3.7.3. Since there is no preferred direction in the (ϕ_1, ϕ_2) plane, variations in the phase of ϕ from one spacetime point to another can be meaningful only if a rule exists for parallel transport through the fibre bundle. In order to attach a meaning to the relative directions of $\phi(x_1)$ and $\phi(x_2)$ in the internal spaces at x_1 and x_2, we need a rule for constructing the wavefunction $\phi(x_1 \to x_2)$ which exists at x_2 and is to count as 'parallel' to $\phi(x_1)$. The physically meaningful change in ϕ between the points x_1 and x_2 is then given by

$$\delta\phi = \phi(x_2) - \phi(x_1 \to x_2). \tag{8.2}$$

Evidently, this is quite similar to the way we defined changes in a vector field in (2.22).

An obvious possibility for such a rule is that the phase angles $\tan^{-1}(\phi_2/\phi_1)$ should be equal for $\phi(x_1 \to x_2)$ and $\phi(x_1)$. It will become apparent that this is the special case corresponding to the absence of electromagnetic fields. Indeed, this rule is equivalent to saying that the ϕ_1 axes at any two points are to count as parallel, and likewise the ϕ_2 axes. In spacetime geometry, the analogous rule, that a single set of self-parallel Cartesian axes can be used to cover the whole manifold, implies that the manifold is flat and that there are no gravitational fields.

A less restrictive rule for parallel transport may be expressed in terms of *connection coefficients* $\Gamma_{ij\mu}$:

$$\phi_i(x \to x + \Delta x) = \phi_i(x) - \Gamma_{ij\mu}(x)\phi_j(x)\Delta x^\mu. \tag{8.3}$$

This rule has the same form as that for parallel transport of spacetime vectors between infinitesimally separated points via the affine connection (2.23), except that the indices i and j refer to directions in the internal space. Unlike the absolute phase, the magnitude of the wavefunction $|\phi| = (\phi^*\phi)^{1/2} = (\phi_1^2 + \phi_2^2)^{1/2}$ has a definite physical meaning in terms of probability amplitudes. We therefore include in the definition of parallel transport the requirement that this magnitude remain unchanged. For this to be so, $\Gamma_{ij\mu}$ must be antisymmetric in i and j (see exercise 8.1) and therefore proportional to the two-dimensional Levi-Civita symbol ϵ_{ij}:

$$\Gamma_{ij\mu}(x) = -\epsilon_{ij}\lambda A_\mu(x). \tag{8.4}$$

The vector field $A_\mu(x)$ will turn out to be essentially the electromagnetic 4-vector potential. The constant λ is intended to allow for different species of particle with different electric charges (proportional to λ). In our present geometrical language, we may say that the wavefunctions representing different particle species exist in independent internal spaces and there is no reason why parallel transport in all these spaces should involve the same connection. If, as we know to be the case, particles of different types respond to the same electromagnetic fields, then we conclude that, as a matter of empirical fact, their connections are determined by the same vector potential A_μ, though possibly with different coefficients λ.

Because we are concerned only with the phases and not the magnitudes of wavefunctions, it is convenient to deal with a fibre bundle whose fibres consist just of these phases. Each fibre can be envisaged as a copy of the unit circle in the complex plane, whose points are labelled by the phase angle θ, with values between 0 and 2π. The angle $\theta(x)$ can be thought of as specifying a transformation. Thus, if we write $\phi(x)$ as $\exp(\mathrm{i}\theta(x))|\phi(x)|$, then the phase factor transforms $|\phi(x)|$ into $\phi(x)$. More generally, a *phase transformation* changes the phase of any complex wavefunction by an angle between 0 and 2π. If the same transformation is made at each spacetime point, then the wave equation satisfied by ϕ is unchanged, and so are all the matrix elements. In this sense, the transformation is a symmetry of the quantum theory. The set of all these transformations constitutes a *symmetry group*. It is possible to consider symmetry groups that are much more general than phase transformations, and the transformations that constitute the group may be labelled by several parameters analogous to θ. The set of all possible values of these parameters is called the *group manifold*. At the most fundamental level, each fibre in a bundle of the kind we are considering is to be thought of as a copy of the group manifold of a symmetry group. In our present case, the symmetry group is called U(1), and its manifold is the unit circle.

In spacetime geometry, we defined objects called tensors by their behaviour under general coordinate transformations. In our fibre bundle, the analogue of a general coordinate transformation is a phase transformation through an angle $\theta(x)$, where $\theta(x)$ is a differentiable function of position. The tensors associated with these transformations are products of ϕ and ϕ^*. For a product $\Phi_{mn} = \phi^{*m}\phi^n$, the transformation law is

$$\Phi'_{mn}(x) = \exp[\mathrm{i}(n-m)\lambda\theta(x)]\Phi_{mn}(x). \tag{8.5}$$

The definition (8.3) of parallel transport leads to a covariant derivative D_μ analogous to (2.24). In terms of the real and imaginary parts of the wavefunction, it is defined by

$$\phi_i(x+\Delta x) - \phi_i(x \to x+\Delta x) = \mathrm{D}_{ij\mu}\phi_j(x)\Delta x^\mu + \mathrm{O}(\Delta x^2) \tag{8.6}$$

which, on account of (8.3), gives

$$\mathrm{D}_{ij\mu}\phi_j(x) = \partial_\mu\phi_i(x) + \Gamma_{ij\mu}(x)\phi_j(x). \tag{8.7}$$

In terms of the complex wavefunction, this may be rewritten as

$$\mathrm{D}_\mu\phi(x) = [\partial_\mu + \mathrm{i}\lambda A_\mu(x)]\phi(x). \tag{8.8}$$

An essential property of a covariant derivative is that it acts on tensors to produce new tensors. In the present context, this means that $\mathrm{D}_\mu\phi$ must have the same phase transformation property as ϕ itself. As for the affine connection in chapter 2, this requirement leads to a transformation law for the connection A_μ which is different from the law (8.5) for 'tensors'. If $\phi' = \exp(\mathrm{i}\lambda\theta)\phi$, then we must have

$$\mathrm{D}'_\mu\phi' = (\partial_\mu + \mathrm{i}\lambda A'_\mu)(\mathrm{e}^{\mathrm{i}\lambda\theta}\phi) = \mathrm{e}^{\mathrm{i}\lambda\theta}\mathrm{D}_\mu\phi = \mathrm{e}^{\mathrm{i}\lambda\theta}(\partial_\mu + \mathrm{i}\lambda A_\mu)\phi \tag{8.9}$$

and so the transformation law for A_μ is

$$A'_\mu(x) = A_\mu(x) - \partial_\mu\theta(x). \tag{8.10}$$

The action of the covariant derivative on a tensor that transforms according to (8.5) must therefore be

$$\mathrm{D}_\mu\Phi_{mn} = [\partial_\mu + \mathrm{i}\lambda(n-m)A_\mu]\Phi_{mn}. \tag{8.11}$$

The set of transformation rules given by (8.5) and (8.10) is usually called a *local gauge transformation* and the connection coefficient A_μ is called a *gauge field*. The derivative D_μ may be called a *gauge-covariant derivative* to distinguish it from the generally covariant derivative ∇_μ.

The intrinsic geometrical structure of spacetime is determined, as we saw in chapter 2, by the metric and by the affine connection. Once the presence of this structure in the dynamical equations of physics has been made explicit through their components $g_{\mu\nu}$ and $\Gamma^\mu_{\nu\sigma}$, we expect that these equations should be generally covariant: that is, their forms should be independent of our choice of a coordinate system. If the dynamical equations are derived from a principle of least action, general covariance is guaranteed by choosing the action to be a scalar.

In the same way, the geometry of the U(1) fibre bundle of electromagnetism (that is, the relationships between phases of wavefunctions at different points in spacetime) is determined by the gauge field $A_\mu(x)$. Once the gauge field has been incorporated into the equations of motion, we expect these equations to be *gauge covariant*. That is, their forms should be preserved by gauge transformations. This will automatically be so if they are derived from a gauge-invariant action. Since we are working in Minkowski spacetime, we shall also require the action to be a Lorentz scalar.

Let us first construct the wave equation for a spin-$\frac{1}{2}$ particle in a prescribed electromagnetic field. In the case of spacetime geometry, an action which is invariant under general coordinate transformations could be built from tensors by contracting all their indices, so that the transformation matrices cancel. Correspondingly, to make a gauge-invariant action, we can use products of gauge tensors, with transformation laws of the form (8.5). Clearly, the product of one such tensor with the complex conjugate of another tensor of the same type will be invariant. Consider the Dirac action (7.61). It should be clear that this will become gauge invariant if we replace the ordinary derivative ∂_μ with the gauge-covariant derivative D_μ:

$$S_{\text{Dirac}} = \int \mathrm{d}^4x\, \bar{\psi}(x)\left(\mathrm{i}\not{\partial} - \lambda \not{A}(x) - m\right)\psi(x). \tag{8.12}$$

The equation that follows from varying $\bar{\psi}$ is

$$\left(\mathrm{i}\not{\partial} - \lambda \not{A}(x) - m\right)\psi(x) = 0. \tag{8.13}$$

This is known as the *minimal coupling prescription*. It is the simplest modification of the Dirac equation that makes it gauge covariant and reduces to the original one when $A_\mu = 0$. A variety of other equations could be invented by introducing further gauge-covariant terms which vanish when $A_\mu = 0$, but there appears to be no good physical reason for doing so.

Some physical consequences of this modified Dirac equation will be explored in the next chapter. A mathematical consequence is that the symmetry of the theory under *global* phase transformations—those which change the phase of the wavefunction by the same amount at each spacetime point—has been promoted to a *local* symmetry, since the phase may be changed by a position-dependent amount $\theta(x)$, provided that a compensating change (8.10) is made in the gauge field. This is precisely analogous to relativistic geometry. In special relativity, Lorentz transformations with a position-independent matrix Λ are global symmetries, and the affine connection coefficients are zero in Cartesian coordinate systems. When the affine connection is explicitly included, general coordinate transformations with position-dependent Λ are symmetries, in the sense of general covariance. In the absence of gravitational fields, coordinate systems may be found in which the connection coefficients are everywhere zero. We shall shortly see that, in the absence of electromagnetic fields, the gauge field can be expressed as the gradient of a scalar function $A_\mu(x) = \partial_\mu\omega(x)$. Therefore, by choosing $\theta(x) = \omega(x) + \text{constant}$ in (8.10), A_μ can be set to zero everywhere. This amounts to choosing a special set of coordinate systems in the fibre bundle, which are analogous to the inertial frames of special relativity.

In addition to (8.13), which describes the response of a charged particle to electromagnetic fields, we need an equation (the analogue of the Einstein field equations) which determines how electromagnetic fields are generated by a distribution of charged particles. The way to find this is again to add to the action a part involving the connection. This must be gauge invariant, and therefore constructed from quantities which are tensors under gauge transformations. The only such quantity that can be built from the gauge field alone is the curvature of the bundle defined, like the Riemann tensor, as the commutator of two covariant derivatives

$$F_{\mu\nu} = -\frac{\mathrm{i}}{\lambda}[\mathrm{D}_\mu, \mathrm{D}_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{8.14}$$

This is in fact gauge invariant, and the simplest Lorentz scalar that can be constructed from it is $F_{\mu\nu}F^{\mu\nu}$. We see that $F_{\mu\nu}$ is none other than the Maxwell field strength tensor given in (3.50) and (3.51). It turns out, as with gravity, that an extremely successful theory is obtained by including in the action only a term proportional to this quantity. This is, essentially, the first term of (3.54). If we identify j_e^μ in (3.54) as proportional to the current density (7.62), then the second term of (3.54) is reproduced by the $A\!\!\!/$ term in (8.12). There is at present no definitive understanding in either theory of why the simplest allowed form of the action should be the one actually selected by nature, although some properties of interacting quantum field theories that we touch on briefly in the next chapter suggest a possible explanation.

To make the correspondence with Maxwell's theory exact, we must examine more closely the role of electric charge. So far, we have established only that the simplest action contains a term *proportional* to $F_{\mu\nu}F^{\mu\nu}$. Allowing for n species of spin-$\frac{1}{2}$ particles, the total action may be written as

$$S = \int \mathrm{d}^4x \left(-\frac{1}{4e^2} F_{\mu\nu}F^{\mu\nu} + \sum_{i=1}^{n} \bar{\psi}_i(x)(\mathrm{i}\partial\!\!\!/ - \lambda_i A\!\!\!/(x) - m_i)\psi_i(x) \right) \tag{8.15}$$

where e is a constant whose value is not known *a priori*. This constant, which may be identified as a fundamental electric charge, is clearly somewhat analogous to the constant G which appears in the theory of gravity. Note, however, that the curvature term in the Einstein–Hilbert gravitational action (4.14) is linear in the Riemann tensor $R_{\mu\nu\sigma\tau}$, which can be contracted to give a non-trivial scalar curvature R. In electromagnetism, the contraction $g^{\mu\nu}F_{\mu\nu}$ is identically zero, because $g^{\mu\nu}$ is symmetric and $F_{\mu\nu}$ is antisymmetric, and the simplest non-trivial Lorentz scalar is quadratic in $F_{\mu\nu}$. This is symptomatic of some important differences between the two theories. The standard form of electromagnetism is obtained by rescaling the fields:

$$A_\mu(x) \to eA_\mu(x) \qquad F_{\mu\nu}(x) \to eF_{\mu\nu}(x) \tag{8.16}$$

after which the action becomes

$$S = \int \mathrm{d}^4x \left(-\frac{1}{4} F_{\mu\nu}F^{\mu\nu} + \sum_{i=1}^{n} \bar{\psi}_i(x)(\mathrm{i}\partial\!\!\!/ - \lambda_i e A\!\!\!/(x) - m_i)\psi_i(x) \right). \tag{8.17}$$

The equations derived from this action describe the electromagnetic interactions of n species of particle with masses m_i and charges $\lambda_i e$. Evidently, the $(n+1)$

constants $(\lambda_1, \ldots, \lambda_n, e)$ are not all independent. We can choose e to be the magnitude of the electronic charge by setting $\lambda_{\text{electron}} = -1$. Then the charges of the remaining particles are multiples, $\lambda_i e$ of this fundamental charge. There is no reason, however, why the λ_i should be integers, or even rational numbers. The fact that the electric charges of all observed particles are integral multiples of a fundamental charge has no explanation within the theory of electromagnetism alone. A possible explanation is offered by the *grand unified theories* of strong, weak and electromagnetic interactions which will be outlined in chapter 12. Notice also that had $F_{\mu\nu}$ contained terms quadratic in A_μ which, as we shall see shortly, is the case in non-Abelian gauge theories, the rescaling of the gauge field in (8.16) would not have removed the charge e entirely from the curvature term, and e would have been a genuine independent parameter.

8.2 Non-Abelian Gauge Theories

The internal spaces in which wavefunctions exist may be more complicated than the complex plane. Consider, for example, the nucleons—the proton and neutron. In processes involving the strong interaction (of which more in chapter 12), they appear on an equal footing: the strong interaction is said to be *charge independent*. This observation, together with the fact that their masses are very similar, leads to the idea that the proton and neutron can be regarded as different states of the same particle—the nucleon. The nucleon wavefunction is then a two-component matrix

$$\psi_{\text{N}}(x) = \begin{pmatrix} \psi_{\text{p}}(x) \\ \psi_{\text{n}}(x) \end{pmatrix}. \tag{8.18}$$

Actually, since the nucleons are spin-$\frac{1}{2}$ particles, each of the two components is itself a four-component spinor, but this does not at present concern us. A state with $\psi_{\text{n}} = 0$ is a pure proton state and *vice versa*, while a state in which both components are non-zero is a superposition of the two. This is quite analogous to the non-relativistic description of spin-$\frac{1}{2}$ polarization states (see appendix B). In particular, any unitary transformation (that is, a rearrangement of the components that leaves the magnitude $(\psi_{\text{N}}^\dagger \psi_{\text{N}})^{1/2}$ unchanged) can be expressed as

$$\psi'_{\text{N}} = \exp[\text{i}(\theta I + \tfrac{1}{2}\boldsymbol{\alpha} \cdot \boldsymbol{\tau})]\psi_{\text{N}} \equiv U(\theta, \boldsymbol{\alpha})\psi_{\text{N}} \tag{8.19}$$

where I is the unit 2×2 matrix, τ^1, τ^2, τ^3 are the Pauli matrices and θ, α^1, α^2, α^3 are real angles. Such transformations are involved, for example, in reactions which change the state of a nucleon but not the total number of nucleons, such as beta decay ($\text{n} \to \text{p} + \text{e}^- + \bar{\nu}_{\text{e}}$) or pion-nucleon scattering ($\pi^- + \text{p} \to \text{n} + \pi^0$). The matrices τ^i have the same numerical values as the spin matrices (7.28), but the symbol $\boldsymbol{\tau}$ emphasizes that they refer to a different internal property of the particles. This property is called *isotopic spin*, or more commonly *isospin*, denoted by $\boldsymbol{T}$. The transformations parametrized by θ are phase transformations, which will not concern us for the moment. The others, of the form

$$U(\boldsymbol{\alpha}) = \exp(\tfrac{1}{2}\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}) \qquad (8.20)$$

can be regarded as rotations in an internal three-dimensional *isospin space*. The proton and neutron states correspond to 'isospin up' and 'isospin down' with respect to a chosen quantization axis in this space.

There now arises a question similar to that which led to electromagnetism. The two-component wavefunction at the spacetime point x is to be thought of as existing in a copy of isospin space erected at x, and we would like to know how the directions of the (T^1, T^2, T^3) axes at different points are related. In contrast to the complex phase, these directions have definite physical meanings, because the proton and neutron are physically identifiable states. Parallel transport of a wave function may be defined by introducing a connection as in (8.3), except that i and j now label the components in (8.18) rather than the real and imaginary parts. If Γ is zero, then a parallelly transported wavefunction that represents, say, a neutron at x also represents a neutron at $x + \Delta x$. If Γ is not zero, then the wavefunction may, after being transported to $x + \Delta x$, represent a superposition of proton and neutron states. Since the connection in (8.3) turned out to be related to the electromagnetic field, we may anticipate that the isospin connection is similarly related to some kind of force field. Evidently, one effect of this force is to turn neutrons into protons, so it might provide a means of describing beta decay.

The right-hand side of (8.3) now corresponds to an infinitesimal rotation of the kind (8.20), so the connection coefficient has the form

$$\Gamma_{ij\mu}(x) = -\tfrac{1}{2}\mathrm{i}A^a_\mu(x)\,(\tau^a)_{ij}. \qquad (8.21)$$

There are three independent gauge fields A^a_μ $(a = 1, 2, 3)$, corresponding to the three independent isospin rotations. This connection acts in the fibre bundle whose typical fibre is the set of all transformations of the form (8.20) or, equivalently, the set of all values of the α^a that lead to distinct transformations. This can be taken as the set of all positive and negative values such that $\boldsymbol{\alpha}\cdot\boldsymbol{\alpha} \leq 4\pi^2$, with the proviso that all values for which the equality holds correspond to the same transformation (see exercise 8.2). This set of transformations constitutes the group SU(2). It is a *non-Abelian* group, which means that two rotation matrices $U(\boldsymbol{\alpha})$ and $U(\boldsymbol{\beta})$ do not commute unless $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ point in the same direction. The group U(1) of electromagnetism is an *Abelian* group, because any two phase transformations commute with each other. One consequence of the non-Abelian nature of isospin rotations is that no arbitrary constant λ appears in the connection (8.21) to distinguish different particle species. This is because, as we shall see in more detail below, the gauge field A^a_μ has an intrinsic scale. For example, a rotation through an angle of π about the T^1 or T^2 axis changes a proton state with $T^3 = \frac{1}{2}$ into a neutron state with $T^3 = -\frac{1}{2}$. The same rotation must produce the same reversal of T^3 when acting on any set of particle wavefunctions that form an isospin multiplet. Therefore, the size of the rotation angles in (8.20) has a definite meaning, common to all particle species, and we have no freedom to introduce an arbitrary parameter as in the Abelian case (8.5). On the other hand, different particle species may fall into isospin multiplets of different sizes. Just as with angular momentum, an isospin-T multiplet has $(2T + 1)$ members. For the moment, the three pions (π^+, π^0, π^-) may serve as an example of an isospin-1 triplet. At present, however, in order to describe the mathematics of non-Abelian theories in its simplest terms, I am not taking proper account of the observed properties of elementary particles. When we come to study the application of these theories to real physical particles, it will be necessary to revise the way in which the particles are assigned to isospin multiplets. The wavefunction for an isospin-T multiplet undergoes parallel transport with a connection similar to (8.21) except that the Pauli matrices are replaced with a suitable set of three $(2T + 1) \times (2T + 1)$ matrices, called the isospin-T *representation* of the group SU(2). The same gauge field appears in each case, however.

Given the gauge connection, we have a gauge-covariant derivative

$$\mathrm{D}_\mu = \partial_\mu + \mathrm{i}A_\mu(x) \tag{8.22}$$

where $A_\mu(x)$ is a matrix defined by

$$A_\mu(x) = A^a_\mu(x)T^a \tag{8.23}$$

and T^a are the isospin matrices appropriate to the particular multiplet of wavefunctions on which the derivative acts. Under a gauge transformation, each multiplet transforms as

$$\psi'(x) = U(\boldsymbol{\alpha})\psi(x) = \exp[\mathrm{i}\boldsymbol{\alpha}(x) \cdot \boldsymbol{T}]\psi(x). \tag{8.24}$$

To find the transformation law for the gauge fields, consider

$$\mathrm{D}'_\mu\psi' = (\partial_\mu + \mathrm{i}A'_\mu)U\psi = \left(U\partial_\mu + \partial_\mu U + \mathrm{i}A'_\mu U\right)\psi. \tag{8.25}$$

The requirement is that this should equal $U\mathrm{D}_\mu\psi$, so A_μ must transform as

$$A'_\mu = UA_\mu U^{-1} + \mathrm{i}(\partial_\mu U)U^{-1}. \tag{8.26}$$

If U were just the phase factor $\exp(\mathrm{i}\theta(x))$, this would be the same as (8.10).

The non-Abelian analogue of the electromagnetic field strength tensor is, naturally, the curvature tensor $-\mathrm{i}[\mathrm{D}_\mu, \mathrm{D}_\nu]$. This is more closely analogous to the Riemann tensor, in the sense that it involves the non-commuting properties of both the derivative ∂_μ and the matrices T^a. It is given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}[A_\mu, A_\nu]. \tag{8.27}$$

Of course, the matrix form of this expression depends on the particular representation of the gauge group (here SU(2)) to which the matrices belong.

However, in every representation, these matrices satisfy the commutation relations of the *Lie algebra*

$$[T^a, T^b] = \mathrm{i}C^{abc}T^c \tag{8.28}$$

where the set of *structure constants* C^{abc} is totally antisymmetric in a, b and c. For SU(2), they are given by $C^{abc} = \epsilon^{abc}$ (see appendix B). The T^a are the generators of the symmetry transformations (in our case, isospin rotations) and in group theory language are called the generators of the symmetry group. Using the definition (8.23) of the matrices A_μ, we find that

$$F_{\mu\nu} = F^a_{\mu\nu}T^a \tag{8.29}$$

where the field strengths

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - C^{abc}A^b_\mu A^c_\nu \tag{8.30}$$

are the same for any representation.

Unlike the electromagnetic field strength tensor, (8.27) is not a gauge-invariant object. In fact, its transformation law is

$$F'_{\mu\nu} = UF_{\mu\nu}U^{-1} \tag{8.31}$$

as readers are invited to verify in exercise 8.3. From this it follows that the three field strengths (8.30) ($a = 1, 2, 3$) belong to an isospin-1 multiplet. To understand this, notice first that (8.31) implies the transformation

$$F^{a\ \prime}_{\mu\nu} = \mathcal{U}^{ab}(\boldsymbol{\alpha})F^b_{\mu\nu} \tag{8.32}$$

where the coefficients $\mathcal{U}^{ab}$ are defined by

$$U(\boldsymbol{\alpha})T^bU^{-1}(\boldsymbol{\alpha}) = \mathcal{U}^{ab}(\boldsymbol{\alpha})T^a. \tag{8.33}$$

It is a group-theoretical fact (which I shall not prove) that, if we regard these coefficients as the elements of a 3 × 3 matrix, it can be written as

$$\mathcal{U}(\boldsymbol{\alpha}) = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{T}}) \tag{8.34}$$

where the 3 × 3 matrices $\mathcal{T}^a$ form a special representation of SU(2) called the *adjoint representation*. Every Lie group possesses such a representation, in which the number of members of the multiplet is equal to the number of independent generators. The matrices of the adjoint representation can be expressed in terms of the structure constants as

$$(\mathcal{T}^a)_{bc} = -\mathrm{i}C^{abc}. \tag{8.35}$$

The proof that these matrices satisfy the commutation relations (8.28) is the subject of exercise 8.4.

Once again, we need to construct a gauge-invariant action for the gauge fields. The simplest possibility is

$$S = -\frac{1}{4g^2}\int \mathrm{d}^4x\, F^a_{\mu\nu}F^{a\mu\nu} \tag{8.36}$$

where g is a *coupling constant* analogous to the electric charge. As in (8.16), we now rescale the gauge field by a factor of g and rename the quantity $g^{-1}F^a_{\mu\nu}$ as $F^a_{\mu\nu}$, to get

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - gC^{abc}A^b_\mu A^c_\nu. \tag{8.37}$$

and

$$S = -\frac{1}{4}\int \mathrm{d}^4x\, F^a_{\mu\nu}F^{a\mu\nu}. \tag{8.38}$$

In the quantum theory, the gauge field becomes a field operator for 'intermediate vector bosons', which mediate the corresponding force. In the case of electromagnetism, these are photons, which are neutral particles. The action (8.38), expressed in terms of the rescaled field strength (8.37), contains products of three As multiplied by g and products of four As multiplied by g^2. It will become clear in the next chapter that such terms represent interactions between the vector bosons of the non-Abelian theory, whose strength is measured by g. Indeed, it is already obvious that the actions for free particles considered in chapters 6 and 7 are only quadratic in the field operators. Thus, these particles carry the 'charge' g of the force which they themselves mediate (in contrast to the photon, which is electrically neutral), and this fact has important physical consequences. The situation is similar in the case of gravity. In order to obtain the wave equation (7.119) for gravitons (which in our present language are 'intermediate tensor bosons' mediating the gravitational force), we had, in effect, to expand the gravitational action in powers of the field $h_{\mu\nu}$, keeping only the quadratic terms. The gravitational analogue of charge is energy density which is, of course, possessed by the gravitons themselves, and the full gravitational action has non-quadratic terms that lead to interactions between gravitons. I should point out, however, that the quantum theory of gravity based on this action appears to be mathematically unsound, for reasons I shall touch on later.

A detail that will be important to us later on concerns the normalization of the generator matrices T^a. The transformation matrix $U = \exp[\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{T}]$ could clearly be written in terms of a new set of matrices, say T'^a, which are linear combinations of the T^a, and a new set of parameters α'^a, such that $\boldsymbol{\alpha}'\cdot\boldsymbol{T}' = \boldsymbol{\alpha}\cdot\boldsymbol{T}$. This would entail a corresponding redefinition of the structure constants C^{abc} and of the gauge fields A^a_μ. Now, these field operators will have the commutation relations that cause them to create and annihilate particle states with the correct orthonormality properties, provided that their action is that shown in (8.38). We should, however, make sure that this action really is gauge invariant. To this end, consider the quantity $\mathrm{Tr}[F_{\mu\nu}F^{\mu\nu}]$. Because of the identity $\mathrm{Tr}[AB] = \mathrm{Tr}[BA]$, valid for any two matrices A and B, this quantity is easily seen to be invariant under the gauge transformation (8.31). We have

$$\mathrm{Tr}[F_{\mu\nu}F^{\mu\nu}] = F^a_{\mu\nu}F^{b\mu\nu}\,\mathrm{Tr}[T^aT^b] \tag{8.39}$$

and this will be proportional to our Lagrangian density $-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}$ provided that the generator matrices satisfy the condition

$$\mathrm{Tr}[T^aT^b] = \lambda\delta^{ab} \tag{8.40}$$

where λ is a constant. Given *some* set of generator matrices, it will always be possible to find linear combinations of them which satisfy this condition, and these will be the ones we use. For our SU(2) theory, the isospin-$\frac{1}{2}$ matrices $T^a = \frac{1}{2}\tau^a$ do satisfy (8.40) with $\lambda = \frac{1}{2}$.

If we include spin-$\frac{1}{2}$ fermions upon which the gauge field acts, the total action is

$$S = \int \mathrm{d}^4x \left(-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} + \sum_{i=1}^{n} \bar{\psi}_i(x)\left(\mathrm{i}\not{\partial} - g\not{A}(x) - m_i\right)\psi_i(x)\right). \tag{8.41}$$

This is now expressed in a rather compact notation. The sum is over multiplets of wavefunctions ψ_i, each having $(2T^{(i)}+1)$ members in the case of SU(2) isospin. Each member is itself a Dirac spinor, so ψ_i may be represented schematically in the form

$$\psi_i = \left.\begin{pmatrix} \left(\vdots\right) \\ \left(\vdots\right) \\ \vdots \\ \left(\vdots\right) \end{pmatrix} \begin{matrix} \}4 \\ \}4 \\ \vdots \\ \}4 \end{matrix}\right\} 2T^{(i)} + 1.$$

The matrix $\not{A}$ is

$$\not{A} = A^a_\mu\gamma^\mu T^{(i)a} \tag{8.42}$$

where $T^{(i)a}$ is the ath generator matrix in the isospin-$T^{(i)}$ representation. The Dirac matrix γ^μ acts on each four-component spinor independently, while $T^{(i)a}$ treats each spinor as a single element.

From the action (8.41) we derive an Euler–Lagrange equation for the gauge field, which is the non-Abelian analogue of Maxwell's equations:

$$\mathrm{D}_\mu F^{\mu\nu} = J^\nu \qquad \text{or} \qquad \partial_\mu F^{a\mu\nu} - gC^{abc}A^b_\mu F^{c\mu\nu} = J^{a\nu}. \tag{8.43}$$

The current is given by

$$J^{a\nu} = g\sum_i \bar{\psi}_i\gamma^\nu T^{(i)a}\psi_i. \tag{8.44}$$

For example, in the case of the nucleon doublet,

$$\begin{aligned} J^{3\nu} &= g(\bar{\psi}_{\rm p}\ \bar{\psi}_{\rm n})\gamma^\nu \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} \psi_{\rm p} \\ \psi_{\rm n} \end{pmatrix} \\ &= g\left[\tfrac{1}{2}\bar{\psi}_{\rm p}\gamma^\nu\psi_{\rm p} - \tfrac{1}{2}\bar{\psi}_{\rm n}\gamma^\nu\psi_{\rm n}\right] \\ &= g\sum_{\rm p,n} T^3 \times (\text{probability current density}). \end{aligned} \tag{8.45}$$

There is, of course, a Dirac equation of the form (8.13) for each multiplet of wavefunctions, λ being replaced by g and $\not{A}$ by (8.42).

We saw in chapter 3 that, as a consequence of gauge invariance, the electromagnetic current $j_{\rm e}^\mu$ is conserved in the classical theory. As readers may easily check using the Dirac equation (8.13), the quantum-mechanical current $j_{\rm e}^\mu = \lambda\bar{\psi}\gamma^\mu\psi$ (which becomes $\lambda e\bar{\psi}\gamma^\mu\psi$ after the rescaling (8.16)) is also conserved. The conservation law $\partial_\mu j_{\rm e}^\mu = 0$ is a gauge-covariant equation because the current is a gauge scalar, with $n = m = 1$ in (8.5), and its gauge-covariant derivative (8.11) is the same as the ordinary derivative. In the non-Abelian theory, however, the current (8.44) is not a gauge scalar. It is a multiplet of currents, whose members are labelled by a, which belongs to the adjoint representation of the gauge group and satisfies the covariant equation

$$\mathrm{D}_\mu J^\mu = 0 \qquad \text{or} \qquad \partial_\mu J^{a\mu} - gC^{abc}A_\mu^b J^{c\mu} = 0. \tag{8.46}$$

The current is said to be *covariantly conserved*, but it clearly is not conserved in the usual sense. This does not, however, imply a breakdown of the general rule that a symmetry implies the existence of a conserved quantity. If we differentiate the non-Abelian Maxwell equation (8.43) and take into account the antisymmetry of $F^{\mu\nu}$, we find that the modified current

$$\tilde{J}^{a\nu} = J^{a\nu} + gC^{abc}A_\mu^b F^{c\mu\nu} \tag{8.47}$$

is conserved in the ordinary sense:

$$\partial_\nu\tilde{J}^{a\nu} = \partial_\nu\partial_\mu F^{a\mu\nu} = 0. \tag{8.48}$$

In fact, $\tilde{J}^{a\nu}$ is the ‘Noether current’ associated with the non-Abelian symmetry. That is, it is the current which ought to be conserved according to Noether’s theorem (see exercise 8.5). The two terms in (8.47) have a simple physical significance. The current represents the flow of isospin, in the same way that the electromagnetic current represents the flow of charge. The first contribution is that of the fermions, and the second is that of the gauge fields or, in the quantized theory, of the vector bosons which, as we have seen, themselves carry isospin.

The components of the field strength tensor (8.37) can be thought of as ‘electric’ and ‘magnetic’ fields $\boldsymbol{E}^a$ and $\boldsymbol{B}^a$. As we saw in chapter 3, (3.45) implies that there are no magnetic monopoles, except at the expense of singularities in the potential $A_\mu(x)$. In the non-Abelian theory, the corresponding equation is

$$\partial_i B^{ai} = gC^{abc} A_i^b B^{ci}. \tag{8.49}$$

Because the right-hand side is non-zero, the non-Abelian theory allows the possibility of 'magnetic monopoles' without singularities in the gauge field. Of course, the non-Abelian 'magnetic field' is not what we ordinarily recognize as a magnetic field. In unified theories, which are more complicated than the ones we have so far discussed, electromagnetism is combined with other forces in a manner which permits the appearance of objects with the properties of genuine magnetic monopoles, and I shall have more to say about this in chapter 13.

8.3 Non-Abelian Theories and Electromagnetism

It is now necessary to understand how the phase transformations of electromagnetism fit in with the SU(2) isospin rotations we have been considering. The general unitary transformation (8.19) includes a phase transformation, which we have so far ignored. Since θ multiplies the unit matrix, any phase transformation commutes with any isospin rotation, so the set of transformations of the form (8.19) constitutes a product group, written as SU(2)×U(1). This means that each transformation is the product of two independent transformations, one from each group. In the transformations considered in the last section, only the identity transformation of U(1) was involved. Now, the U(1) component of this product group cannot correspond directly to electromagnetism because it changes the phase of the electrically charged proton and the neutral neutron by the same amount. To represent electromagnetism in this context, we must look for transformations of the form (8.19) in which the angles θ and $\boldsymbol{\alpha}$ are related in such a way that the net transformation changes the phase of ψ_{p} while leaving the phase of ψ_{n} unchanged. The relation that achieves this is

$$\theta = \tfrac{1}{2}Y\omega \qquad \alpha_1 = \alpha_2 = 0 \qquad \alpha_3 = \omega \tag{8.50}$$

where ω is an arbitrary angle and Y is a constant, which in this case is $Y = 1$. With this relation, we have

$$\theta I + \tfrac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau} = \omega\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} \tag{8.51}$$

and

$$U(\theta, \boldsymbol{\alpha}) = \begin{pmatrix}\mathrm{e}^{\mathrm{i}\omega} & 0\\ 0 & 1\end{pmatrix} \tag{8.52}$$

which is the desired transformation matrix. Since any two matrices of the form (8.52) commute with each other, the set of all such transformations is a U(1) subgroup of SU(2)×U(1), and quite suitable for representing electromagnetism.

If this scheme is to work, it must be possible to assign to each isospin multiplet a value of Y, called *hypercharge*, in such a way that the matrix corresponding to (8.52) correctly reflects the charges of all the particles in the multiplet. That is, if we use in (8.51) the isospin matrices appropriate for the particular multiplet, the transformation matrix must turn out to have the form

$$U(\theta, \boldsymbol{\alpha}) = \begin{pmatrix} e^{iQ_1\omega} & 0 & \cdots \\ 0 & e^{iQ_2\omega} & \\ \vdots & & \ddots \end{pmatrix} \tag{8.53}$$

where the Q_i are the charges of the particles in the multiplet, measured as multiples of a fundamental charge. This will be so if the charges are related to the T^3 quantum numbers of the particles by

$$Q = T^3 + \tfrac{1}{2}Y. \tag{8.54}$$

It so happens that relations of just this kind, the *Gell-Mann–Nishijima relations*, are needed for the phenomenological classification of the observed particles. For example, $Y = 1$ and $T^3 = \pm\frac{1}{2}$ for the nucleon doublet and $Y = 0$, $T^3 = (1, 0, -1)$ for the pions.

8.4 Relevance of Non-Abelian Theories to Physics

Had we not already known of the existence of electromagnetic forces, the geometrical considerations of §8.1 might have led us to predict the occurrence of such forces in nature. Can we, then, identify forces in nature that correspond to the extension of these geometrical ideas to non-Abelian symmetry groups? The answer to this is a qualified ‘yes’. The idea of non-Abelian gauge theories was first suggested by C. N. Yang and R. L. Mills in 1954, and theories of this kind are generally known as *Yang–Mills theories*. At that time, it appeared that observed particles such as protons, neutrons and pions were truly fundamental, and the theory of Yang and Mills was based on the approximate nuclear isospin symmetry which relates these particle states in the way I have described. It is now believed that the nucleons, pions and other strongly-interacting particles are themselves composed of more fundamental particles, the *quarks*. The experimental evidence for this, although compelling, is indirect. It appears that quarks are permanently bound inside the observed particles, and no quark has ever been detected in isolation. The nuclear isospin symmetry, part of what is now known as *flavour* symmetry, appears to be more or less accidental and the proton and neutron, for example, are not to be regarded as different states of the same particle in the straightforward way suggested by (8.18). However, it is consistent with our present knowledge to group the quarks, and also the *leptons*, which include the electron, muon and tau particle, together with their associated neutrinos, into multiplets of a different symmetry called *weak isospin*. This is also an SU(2) symmetry and can be combined, as above, with phase transformations to give SU(2)×U(1).

The gauge theory associated with this symmetry can be identified as describing the electromagnetic and weak interactions. As it happens, the proton and neutron can loosely be considered as forming a weak isospin doublet, in the sense that converting a proton into a neutron involves changing one of its constituent quarks, called an 'up' (u) quark into a 'down' (d) quark, and these two quarks form a weak isospin doublet. Therefore, the picture of beta decay as parallel transport in the presence of a non-trivial gauge connection survives in this version of the theory. Quantum-mechanically, what happens is that a d quark in a neutron, say, turns into a u quark by emitting a gauge quantum, a particle called W^-, whose field operator is one of the gauge fields, which then decays into an electron and an antineutrino.

To construct a theory of such processes, which I shall describe more thoroughly in chapter 12, an important obstacle must be overcome. Unlike electromagnetic forces, the weak interaction which is responsible for beta decay has a very short range. As will become clear in the next chapter, this implies that the gauge quanta must have rather large masses. In fact, the W^- is observed to be about 100 times as massive as the proton. Since its field is a 4-vector, it is a spin-1 particle, whose wave equation is the Proca equation (7.110), and it is easy to see that the mass term in this equation originates with a term $\frac{1}{2}A_\mu^a A^{a\mu}$ in the Lagrangian density. No such term appears in (8.41), for the very good reason that it is not gauge invariant. In order to interpret the SU(2)×U(1) theory in terms of electroweak interactions, therefore, we have to understand how massive gauge quanta can emerge from a gauge-invariant theory. This requires the idea of *spontaneous symmetry breaking*, which will be introduced in chapter 11.

8.5 The Theory of Kaluza and Klein

Now that we have seen how theories of electromagnetism and other forces arise from much the same sort of geometrical considerations as the relativistic theory of gravity, it is natural to wonder whether the analogy can be made any more concrete. In other words, are the origins of gravity and other forces not merely similar but identical? T Kaluza (1921) and O Klein (1926) put forward a theory in which gravity and electromagnetism appear as two different aspects of exactly the same phenomenon. According to this theory, the vector potential A_μ is part of the metric tensor of a five-dimensional spacetime.

Setting aside, temporarily, the fact that we perceive only four dimensions, let us call the five-dimensional metric tensor $\widetilde{g}_{AB}$. To emphasize the extra dimension, I shall let the indices A and B take the values 0, 1, 2, 3, 5. We redefine the components of $\widetilde{g}_{AB}$ as follows:

$$\widetilde{g}_{5\mu} = \widetilde{g}_{\mu 5} = \widetilde{g}_{55} A_\mu \qquad \widetilde{g}_{\mu\nu} = g_{\mu\nu} + \widetilde{g}_{55} A_\mu A_\nu \tag{8.55}$$

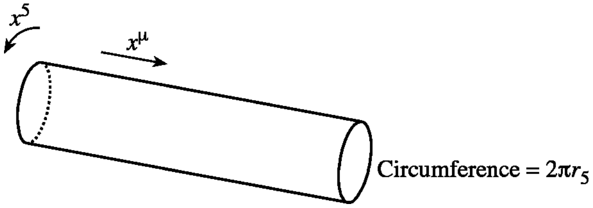


Figure 8.1. Two-dimensional representation of the five-dimensional Kaluza–Klein spacetime.

where the indices μ and ν run from 0 to 3 as usual. The action for five-dimensional gravity is

$$S = -\frac{1}{16\pi\widetilde{G}}\int \mathrm{d}^5x\, \widetilde{g}^{1/2}\widetilde{R} \tag{8.56}$$

where the gravitational constant $\widetilde{G}$, the metric determinant $\widetilde{g}$ and the curvature scalar $\widetilde{R}$ are the five-dimensional ones. If we take the extra dimension to be spacelike, then $\widetilde{g}_{55}$ is negative and $\widetilde{g}$ is positive. We now make two assumptions:

(i) $g_{\mu\nu}$ and A_μ are independent of x^5 and g_{55} is just a constant;

(ii) the five-dimensional spacetime manifold has the structure illustrated in figure 8.1. In the fifth dimension it is of finite extent and closes to form a cylinder of radius r_5.

To account for the unobservability of the fifth dimension, we simply take r_5 to be much smaller than any length scale on which measurements can be made.

If, using these assumptions, (8.55) is substituted into (8.56), the result is

$$S = -\int \mathrm{d}^4x\,(-g)^{1/2}\left(\frac{1}{16\pi G}R + \frac{1}{4e^2}F_{\mu\nu}F^{\mu\nu}\right) \tag{8.57}$$

where g and R are the four-dimensional quantities, and $F_{\mu\nu}$ is the Maxwell field strength tensor. (In principle, partial derivatives $\partial_\mu A_\nu$ are replaced with covariant ones, $\nabla_\mu A_\nu$, but in fact the affine connection terms cancel from $F_{\mu\nu}$.) This is precisely the action we need to describe a spacetime in which both gravitational and electromagnetic fields are present. The four-dimensional gravitational constant G and the charge e are given in terms of the original parameters by

$$G = \widetilde{G}/2\pi r_5|\widetilde{g}_{55}|^{1/2} \qquad \text{and} \qquad e^2 = 8\widetilde{G}/r_5|\widetilde{g}_{55}|^{3/2}. \tag{8.58}$$

Readers may like to be warned that this simple and natural-looking result is quite complicated to verify. Thus, we use (2.50) to work out the five-dimensional affine connection coefficients, separating out those which have only μ indices from those which have one or more indices equal to 5. We substitute the result into (2.36) to get the five-dimensional Ricci tensor and contract this with the five-dimensional metric tensor to get $\widetilde{R}$. That the result of all this boils down to (8.57) strikes me as a minor miracle!

Appealing though this theory is, little attention was paid to it for a long time. Partly, no doubt, this was because it leads to no new observable effects. An unsatisfactory feature is that the two assumptions needed to obtain the final result have no particular justification. The theory would be greatly improved if some dynamical explanation could be found: that is, if it could be shown that a more general five-dimensional metric would naturally evolve into one approximately described by (8.57). Unfortunately, no such mechanism is known. It is worth mentioning that assumption (i) can be relaxed by expanding $g_{\mu\nu}$ and A_μ as Fourier series in x^5. For the reasons indicated in exercise 8.6, the additional terms correspond to wavefunctions or field operators for particles with very large masses, which we would not expect to have observed. In this sense, assumption (i) can be regarded as a natural consequence of assumption (ii).

More complicated non-Abelian gauge theories can be obtained in much the same way, by starting with more dimensions and *compactifying* them in various ways. In recent years, the Kaluza–Klein idea has been much studied because a number of theories, the *supergravity* and *superstring* theories, either can be more simply formulated in more than four dimensions or are mathematically consistent only in some number of dimensions greater than four. The simpler aspects of some of these theories will be explored in chapter 15.

Exercises

8.1. If the real and imaginary parts of ϕ are changed to $\phi_i + \delta\phi_i$, what is the first-order change in the magnitude of ϕ? Show that parallel transport using the connection coefficients (8.4) leaves the magnitude of ϕ unchanged.

8.2. In the transformation matrix (8.20), let $\boldsymbol{\alpha} = \alpha\boldsymbol{n}$, where $\boldsymbol{n}$ is a unit vector. Show that $(\boldsymbol{\tau}\cdot\boldsymbol{n})^2 = 1$ and that

$$\exp(\mathrm{i}\alpha\boldsymbol{\tau}\cdot\boldsymbol{n}/2) = \cos(\alpha/2) + \mathrm{i}\sin(\alpha/2)(\boldsymbol{\tau}\cdot\boldsymbol{n}).$$

Show that an angle $\alpha + 4\pi$ leads to the same transformation as α and that all distinct transformations are included if α is restricted to the range $-2\pi \le \alpha \le 2\pi$. Hence show that the range of values of $\boldsymbol{\alpha}$ which all correspond to distinct transformations is $\boldsymbol{\alpha}\cdot\boldsymbol{\alpha} \le 4\pi^2$, except that all values of $\boldsymbol{\alpha}$ for which the equality holds correspond to $U = -1$.

8.3. A matrix U and its inverse U^{-1} are related by $UU^{-1} = I$. Show that, if U depends on x, then $\partial_\mu U^{-1} = -U^{-1}(\partial_\mu U)U^{-1}$. For the gauge-transformed field (8.26), show that

$$\partial_\mu A'_\nu = U\Big\{\partial_\mu A_\nu + [U^{-1}\partial_\mu U, A_\nu] \\ +\mathrm{i}U^{-1}(\partial_\mu\partial_\nu U) - \mathrm{i}U^{-1}(\partial_\nu U)U^{-1}(\partial_\mu U)\Big\}U^{-1}.$$

Hence verify (8.31)

8.4. For any three matrices T^a, T^b and T^c, verify the *Jacobi identity*

$$[[T^a, T^b], T^c] + [[T^b, T^c], T^a] + [[T^c, T^a], T^b] = 0.$$

Taking these matrices to obey the Lie algebra relations (8.28), show that the structure constants C^{abc} satisfy

$$C^{abd}C^{dce} + C^{bcd}C^{dae} + C^{cad}C^{dbe} = 0.$$

Hence show that the matrices defined by (8.35) obey (8.28).

8.5. (a) Consider a field theory containing a collection of field components $\{\phi_i(x)\}$. The index i labels *all* the components of *all* the fields, which may include both bosons and fermions. (In the case of a gauge field A^a_μ, for example, i includes both a and μ.) The Lagrangian density can be expressed as a function of these field components and their spacetime derivatives, $\mathcal{L}(\{\phi_i\}, \{\partial_\mu\phi_i\})$. Show that the Euler–Lagrange equations are

$$\partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi_i)}\right) = \frac{\partial\mathcal{L}}{\partial\phi_i}.$$

(b) Suppose that $\mathcal{L}$ has a symmetry, such that it is unchanged to first order in a set of small parameters ϵ^a when the fields undergo the infinitesimal changes

$$\phi_i \to \phi_i + \epsilon^a f_i^a(\phi) \qquad \partial_\mu\phi_i \to \partial_\mu\phi_i + \epsilon^a\partial_\mu f_i^a(\phi).$$

Generalize the considerations of §3.2 to prove the field-theoretic version of Noether's theorem, which asserts that the current

$$j^{a\mu}(x) = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi_i)} f_i^a(\phi)$$

is conserved ($\partial_\mu j^{a\mu} = 0$). As usual, a sum over the repeated index i is implied.
(c) Consider the special case of the gauge transformations (8.24) and (8.26) for which the angles α^a are infinitesimal and independent of x. Show that the infinitesimal transformations in the fields are

$$\psi_i \to \psi_i + \mathrm{i}\alpha^a T^a\psi_i \qquad A^b_\nu \to A^b_\nu + \alpha^a C^{abc} A^c_\nu$$

and verify that the corresponding conserved current is proportional to that given in (8.47).

8.6. Show that the five-dimensional Kaluza–Klein metric $\tilde{g}_{AB}$ can be written in the form

$$\tilde{g}_{AB} = \begin{pmatrix} I & (\tilde{g}_{55})^{1/2}A_\mu \\ 0 & (\tilde{g}_{55})^{1/2} \end{pmatrix} \begin{pmatrix} g_{\mu\nu} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I & 0 \\ (\tilde{g}_{55})^{1/2}A_\nu & (\tilde{g}_{55})^{1/2} \end{pmatrix}.$$

The elements of each matrix represent, in clockwise order from the top left, a 4×4 matrix, a four-component column, a single element, and a four-component row. $g_{\mu\nu}$ is the four-dimensional metric and I the 4×4 unit matrix. Hence show that the five-dimensional inverse matrix $\widetilde{g}^{AB}$ has elements $\widetilde{g}^{\mu\nu} = g^{\mu\nu}$, $\widetilde{g}^{5\mu} = \widetilde{g}^{\mu 5} = -A^{\mu}$ and $\widetilde{g}^{55} = A_{\mu}A^{\mu} + (\widetilde{g}_{55})^{-1}$, and that the five-dimensional metric determinant is $\det(\widetilde{g}_{AB}) = \widetilde{g}_{55} \det(g_{\mu\nu})$.

Consider a scalar field with the five-dimensional action

$$S = \int \mathrm{d}^5 x \, \widetilde{g}^{1/2} \, \widetilde{g}^{AB} \, \partial_A \widetilde{\phi}^* \partial_B \widetilde{\phi}.$$

Assume that $\widetilde{\phi}(x, x^5) = \exp(\mathrm{i}\lambda x^5)\phi(x)$, where x denotes the four-dimensional coordinates. When the extra dimension is compactified, show that $\phi(x)$ can be interpreted as the field for particles with charge λe and a mass given by $m^2 = -\lambda^2/\widetilde{g}_{55}$. Given that $\widetilde{\phi}$ should be a single-valued function of x^5, what values of λ are permissible?

Chapter 9

Interacting Relativistic Field Theories

The relativistic wave equations and field theories encountered in chapter 7 described only the properties of free, non-interacting particles. The wave equation for a free particle is always of the form (differential operator)$\phi = 0$, and therefore the corresponding Lagrangians are quadratic in the fields. We have already seen that gauge theories give rise, in a natural way, to Lagrangians that contain terms of higher than quadratic order in the fields, and these terms describe interactions. In (8.41), for example, $\bar{\psi}\not{A}\psi$ describes an interaction between a fermion and a gauge field, while the higher-order terms in $F^a_{\mu\nu}F^{a\mu\nu}$ describe interactions of the gauge fields amongst themselves. It is, of course, only in the presence of interactions that physically interesting events can occur. At the same time, the physical interpretation of interacting quantum field theories is rather difficult. The interpretation of free field theories is based on expansions such as (7.80) in terms of solutions of the appropriate wave equation, the coefficients being interpreted as creation and annihilation operators. When a fermion interacts with a gauge field, the Dirac equation is modified as in (8.13). If the gauge field is itself an operator, this equation cannot be solved for ψ alone, and the plane-wave solutions of the free theory have no definite significance. It is, of course, possible to write the field as a Fourier transform, but the momentum k^μ no longer satisfies the constraint $k_\mu k^\mu = m^2$. Although field operators still have the canonical commutation relations, such as (7.87) for Dirac spinors, the coefficients in the Fourier expansion no longer have the simple commutation relations required for creation and annihilation operators.

To make sense of interacting theories, it is generally advantageous to have in mind some particular kind of experiment whose outcome we want to predict. More often than not, the experiments to which relativistic field theory is relevant are high-energy scattering experiments. These are, indeed, the most fruitful method of probing the fundamental structure of matter, and it is with a view to interpreting such experiments that much of the mathematics of interacting field theories has been developed. I shall begin, therefore, by discussing the field-theoretic aspects of this interpretation.

9.1 Asymptotic States and the Scattering Operator

The multi-particle states encountered in free field theories are eigenstates of the Hamiltonian, so they can exist unchanged for as long as the system is left undisturbed. In an interacting theory, the eigenstates of the Hamiltonian cannot, in general, be characterized by a definite number of particles with definite energies and momenta. Indeed, it is not often possible to discover exactly what these eigenstates are. It is reasonable to suppose that the ground state is recognizable as the vacuum. Another reasonable assumption is that a single, stable particle can exist in isolation for an indefinite time, so that these single-particle states would also be energy eigenstates. If the second assumption is valid, it might appear that each stable particle would be represented in the theory by a field operator which creates it from the vacuum and, conversely, that each field operator in the theory could act on the vacuum to create a stable, single-particle state. This, however, is not so. For example, the standard model of particle physics described in chapter 12 contains, amongst others, field operators for quarks and muons. While muons are indeed observed experimentally, they eventually decay (with a lifetime of about 2×10^{-6} s) into electrons and neutrinos, so a single-muon state cannot be a true energy eigenstate. Quarks, on the other hand, are never observed in isolation, so a single-quark state is not even approximately an eigenstate of the Hamiltonian. Within the standard model, the proton is a true eigenstate, but the operator that creates it from the vacuum is a complicated combination of quark and other field operators. (This statement is believed to be true, being consistent with observations and approximate calculations, but it has not, as far as I know, been rigorously proved.) According to grand unified theories, protons can also decay into lighter particles, and so even the proton is not an eigenstate. At the time of writing, however, proton decay has not been observed.

A second difficulty of interpretation is that, although single particles have, within experimental resolution, well-defined energies and momenta, they also follow quite well-defined paths (seen, for example, as narrow tracks in cloud chamber photographs) and so cannot, strictly speaking, be described by plane waves. This is not a difficulty of principle, because it is quite possible to represent these particles by localized wave packets, whose spread in momentum is well within the range allowed by experimental resolution. Such wave packets are, however, inconvenient to deal with. The standard formalism of interacting field theories is based on a compromise between the strict mathematics and the need for a straightforward interpretation of actual observations. The arguments I am about to present are not really adequate for problems such as the confinement of quarks inside hadrons, but the necessary modifications can be introduced at a later stage.

The processes in which particles scatter or decay are always observed to occur within a very small spacetime region, called the *interaction region*. Outside the interaction region, particles behave, to an extremely good approximation, as if they were free. The initial and final multi-particle states can therefore be approximated as eigenstates of the Hamiltonian of a non-interacting theory. The real reason for this is that particle wavefunctions outside the interaction region are wave packets which do not overlap appreciably. It is convenient to imagine, however, that the interactions are 'switched off' at times well before and after the scattering or decay event takes place. This should be allowable, since the interactions have no significant effect at these times. I shall denote all the field operators collectively by ϕ (dropping the ˆ for simplicity of notation), and the free-particle Hamiltonian by $H_0(\phi)$. Then, taking the event to occur at around $t = 0$, we replace the true Hamiltonian by

$$H(\phi) = H_0(\phi) + \mathrm{e}^{-\epsilon|t|}H_{\mathrm{int}}(\phi) \tag{9.1}$$

where H_{int} is the part of the Hamiltonian that contains interactions and ϵ is a small parameter, which will be set to zero at a suitable stage of the calculation. The modified Hamiltonian reduces to H_0 at $t = \pm\infty$, but if ϵ is small enough, it is essentially the same as the true Hamiltonian within the interaction region. This mathematical device is known as *adiabatic switching*. At very early or very late times, referred to as the 'in' and 'out' region respectively, we no longer need localized wave packets to prevent the particles from interacting, and the wavefunctions of the incoming and outgoing particles can be taken as plane waves.

The field operators $\phi(\boldsymbol{x}, t)$ are, of course, Heisenberg-picture operators, whose evolution with time depends on the Hamiltonian. In the 'in' and 'out' regions, they should behave approximately as free fields. We therefore assume that

$$\begin{aligned} \phi(\boldsymbol{x}, t) &\approx Z^{1/2}\phi_{\mathrm{in}}(\boldsymbol{x}, t) \quad && \text{for } t \to -\infty \\ &\approx Z^{1/2}\phi_{\mathrm{out}}(\boldsymbol{x}, t) \quad && \text{for } t \to +\infty \end{aligned} \tag{9.2}$$

where ϕ_{in} and ϕ_{out} are free field operators and Z is a constant, called the *wavefunction renormalization constant*, which allows the magnitude of the 'in' and 'out' fields to be adjusted in accordance with the correct normalization of the states they create. (Close inspection reveals that some care is needed in interpreting (9.2), but I must refer readers to the more specialized literature for a discussion of this point.) Unlike the interacting fields, the 'in' and 'out' fields can be expanded in terms of plane-wave solutions of the appropriate wave equations, the coefficients being interpreted as particle creation and annihilation operators. The initial state of particles about to undergo scattering will be of the form

$$|k_1, \ldots, k_N; \mathrm{in}\rangle = a_{\mathrm{in}}^{\dagger}(k_N)\cdots a_{\mathrm{in}}^{\dagger}(k_1)|0\rangle. \tag{9.3}$$

In most cases, N will be 1 for a decaying particle or 2 for a pair of colliding particles. The creation operators will be those appropriate for the particular particle species involved. Possible final states, or 'out' states, may be constructed in the same way using 'out' operators. The 'in' and 'out' states are known collectively as *asymptotic states*.

The ‘in’ states are eigenstates of the Hamiltonian $H_0(\phi_{\text{in}})$, but not of the true Hamiltonian $H(\phi)$ which governs the actual time evolution. In the Heisenberg picture, a state vector such as (9.3) stands for the whole history of the system, but its meaning depends on the Hamiltonian. Thus, (9.3) stands for that state which, in the remote past, consisted of N particles with momenta $k_1, \ldots, k_N$, but this does not mean that the state will continue to consist of these N particles. The analogously defined ‘out’ state stands for that state which, in the remote future, will consist of .... Thus, the probability amplitude to detect final state particles with momenta $k'_1, \ldots, k'_{N'}$ given the initial state (9.3) is

$$\langle k'_1, \ldots, k'_{N'}; \text{out}|k_1, \ldots, k_N; \text{in}\rangle \tag{9.4}$$

and one of the primary tasks of field theory is to calculate these amplitudes. The important but mundane process of converting these amplitudes into directly measurable quantities such as decay rates and scattering cross-sections is discussed in appendix D. It is reasonable to assume that the same multi-particle states can exist in the ‘out’ region as in the ‘in’ region, and so there should be a one-to-one correspondence between ‘in’ states and ‘out’ states. This correspondence is expressed in terms of the *scattering operator* S:

$$|k_1, \ldots, k_N; \text{in}\rangle = S|k_1, \ldots, k_N; \text{out}\rangle. \tag{9.5}$$

Thus, the amplitude (9.4) can be expressed as a matrix element of S between two ‘in’ states and is called an *S-matrix element*. To preserve the normalization of the asymptotic states, thereby ensuring that the total probability of a given initial state evolving into *some* final state is 1, the operator S must be unitary. It follows that $\langle \ldots; \text{out}| = \langle \ldots; \text{in}|S$.

9.2 Reduction Formulae

The S-matrix elements can be expressed in terms of the field operators of the interacting theory by means of the *LSZ reduction formula*, named after its inventors H Lehmann, K Symanzik and W Zimmerman. I shall derive an example of such a formula for the case of a single scalar field. The creation and annihilation operators for particles in the ‘in’ and ‘out’ regions can be expressed in terms of the ‘in’ and ‘out’ fields through (7.12) and (7.13). We now apply the identity

$$\int_{-\infty}^{\infty} \mathrm{d}t \, \frac{\partial f(t)}{\partial t} = \lim_{t\to\infty} f(t) - \lim_{t\to-\infty} f(t) \tag{9.6}$$

and the assumed limits (9.2) to write

$$\begin{aligned} a_{\text{in}}(k) - a_{\text{out}}(k) &= \left[\lim_{t\to-\infty} - \lim_{t\to\infty}\right] \mathrm{i}Z^{-1/2} \int \mathrm{d}^3x \, \mathrm{e}^{\mathrm{i}k\cdot x} \overleftrightarrow{\partial_0} \phi(x) \\ &= -\mathrm{i}Z^{-1/2} \int_{-\infty}^{\infty} \mathrm{d}x^0 \, \partial_0 \left(\int \mathrm{d}^3x \, \mathrm{e}^{\mathrm{i}k\cdot x} \overleftrightarrow{\partial_0} \phi(x) \right). \end{aligned} \tag{9.7}$$

If we use the fact that $k^2 = m^2$ and integrate by parts, ignoring any surface term, we can rewrite this as

$$a_{\text{in}}(k) - a_{\text{out}}(k) = -\mathrm{i}Z^{-1/2}\int \mathrm{d}^4x\, \mathrm{e}^{\mathrm{i}k\cdot x}(\Box + m^2)\phi(x). \tag{9.8}$$

Let us use this result to find an expression for the probability amplitude $\langle k'; \text{out}|k; \text{in}\rangle$ for a particle of momentum k' to be found in the distant future, given a single-particle state of momentum k in the distant past. The first step is to write $\langle k'; \text{out}|$ as $\langle 0|a_{\text{out}}(k')$ and re-express $a_{\text{out}}(k')$ using (9.8). The action of $a_{\text{in}}(k')$ on $|k; \text{in}\rangle$ is given by (6.10), but with a relativistic normalization factor as in (7.17) and (7.18), so we get

$$\begin{aligned}\langle k'; \text{out}|k; \text{in}\rangle = {} & (2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k} - \boldsymbol{k}') \\ & + \mathrm{i}Z^{-1/2}\int \mathrm{d}^4x\, \mathrm{e}^{\mathrm{i}k'\cdot x}(\Box + m^2)\langle 0|\phi(x)|k; \text{in}\rangle.\end{aligned} \tag{9.9}$$

Now, we want to use the same method to create $|k; \text{in}\rangle$ from the vacuum. Obviously, we have

$$\langle 0|\phi(x)|k; \text{in}\rangle = \langle 0|\phi(x)a_{\text{in}}^{\dagger}(k)|0\rangle \tag{9.10}$$

but by using (9.8) directly we would get an unwanted term $\langle 0|\phi a_{\text{out}}^{\dagger}|0\rangle$. If, instead, we could arrange to get $\langle 0|a_{\text{out}}^{\dagger}\phi|0\rangle$, then this term could be eliminated, because $\langle 0|a_{\text{out}}^{\dagger} = (a_{\text{out}}|0\rangle)^{\dagger} = 0$. To this end, remember that a_{in} and a_{out} arise from the limits $t \to -\infty$ and $t \to \infty$ respectively in the time integral in (9.8). Therefore, we can arrange the desired ordering of operators by defining the *time-ordered product*

$$\begin{aligned} T[\phi(x)\phi^{\dagger}(y)] &= \phi(x)\phi^{\dagger}(y) \quad \text{if } x^0 > y^0 \\ &= \phi^{\dagger}(y)\phi(x) \quad \text{if } y^0 > x^0 \end{aligned} \tag{9.11}$$

in which the operator referring to the latest time stands on the left. Then, using the adjoint of (9.8), we find

$$\begin{aligned} & \mathrm{i}Z^{-1/2}\int \mathrm{d}^4y\, \mathrm{e}^{-\mathrm{i}k\cdot y}(\Box_y + m^2)\langle 0|T[\phi(x)\phi^{\dagger}(y)]|0\rangle \\ & \qquad = \langle 0|\phi(x)a_{\text{in}}^{\dagger}(k)|0\rangle - \langle 0|a_{\text{out}}^{\dagger}(k)\phi(x)|0\rangle \end{aligned} \tag{9.12}$$

and the last term vanishes. Finally, we substitute this into (9.9) to obtain the reduction formula

$$\begin{aligned}\langle k'; \text{out}|k; \text{in}\rangle = {} & (2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k} - \boldsymbol{k}') + (\mathrm{i}Z^{-1/2})^2 \\ & \times \int \mathrm{d}^4x\, \mathrm{d}^4y\, \mathrm{e}^{\mathrm{i}(k'\cdot x - k\cdot y)}(\Box_x + m^2)(\Box_y + m^2)\langle 0|T[\phi(x)\phi^{\dagger}(y)]|0\rangle.\end{aligned} \tag{9.13}$$

The S-matrix element has now been expressed entirely in terms of the original interacting field, so at this point we can take $\epsilon = 0$ in (9.1) and forget about the 'in' and 'out' fields.

The quantity $-\mathrm{i}\langle 0|T[\phi(x)\phi^{\dagger}(y)]|0\rangle$ is called the *Feynman propagator* for the field ϕ. If translational invariance holds, in both space and time, then it depends only on $(x-y)$ and may be written as a Fourier transform

$$G_{\mathrm{F}}(x-y) = -\mathrm{i}\langle 0|T[\phi(x)\phi^{\dagger}(y)]|0\rangle = \int \frac{\mathrm{d}^4 x}{(2\pi)^4}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\widetilde{G}_{\mathrm{F}}(k). \qquad (9.14)$$

If we use this Fourier transform in the reduction formula and integrate by parts to let the derivatives act on the exponential, we get

$$\begin{aligned}\langle k'; \mathrm{out}|k; \mathrm{in}\rangle = {} & (2\pi)^3 2\omega(\boldsymbol{k})\delta^3(\boldsymbol{k}-\boldsymbol{k}') \\ & + \mathrm{i}(\mathrm{i}Z^{-1/2})^2(2\pi)^4\delta^4(k-k')(k^2-m^2)^2\widetilde{G}_{\mathrm{F}}(k). \qquad (9.15)\end{aligned}$$

Since k and k' are the 4-momenta of free particles, they satisfy $(k^2 - m^2) = (k'^2 - m^2) = 0$. Therefore, the second term is zero unless $\widetilde{G}_{\mathrm{F}}(k)$ has a singularity at $k^2 = m^2$. The form of the propagator depends on the nature of the interactions. If they are such that the particles created by ϕ are stable, then $\widetilde{G}_{\mathrm{F}}(k)$ will turn out to behave roughly as $(k^2 - m^2)^{-1}$. The second term in (9.15) is then zero. In that case, the *single-particle* 'in' and 'out' states satisfy the same orthogonality relation (7.18) as in a free field theory. This means that $|k; \mathrm{in}\rangle$ and $|k; \mathrm{out}\rangle$ are the same state, as we would expect for a single stable particle. If, on the other hand, the ϕ particles can decay into lighter ones, it will turn out that $\widetilde{G}_{\mathrm{F}}(k)$ is roughly of the form $(k^2 - m^2)^{-2}\Gamma(k)$, where $\Gamma(k)$ is related to the probability per unit time for the decay process to occur. In that case, (9.15) can roughly be interpreted as the statement (probability of survival) = 1−(probability of decay). The set of 4-momenta which satisfy $k^2 = m^2$ is called the *mass shell*. Quantities like the propagator, known generically as *Green functions*, are well defined for more general 4-momenta, but S-matrix elements such as (9.15) involve only the *residues* of poles of these Green functions at $k^2 = m^2$: the *on-shell* residues.

It should be clear that the operations which led to the reduction formula (9.13) can be repeated for initial and final states that contain more than one particle. Thus, all S-matrix elements can be expressed in terms of vacuum expectation values of time-ordered products of field operators, $\langle 0|T[\phi(x_1)\cdots\phi^{\dagger}(x_N)]|0\rangle$, where N is the total number of incoming and outgoing particles. The T product orders all the operators according to their time arguments, with the latest on the left and the earliest on the right. When spin-$\frac{1}{2}$ particles are involved, the exponentials in (9.13) are replaced by plane-wave solutions of the Dirac equation and the Klein–Gordon operator $(\Box + m^2)$ by the Dirac operator $(\mathrm{i}\not{\partial} - m)$. Thus, for single particles, (9.13) becomes

$$\langle k', s'; \mathrm{out}|k, s; \mathrm{in}\rangle = (2\pi)^3 2\omega(\boldsymbol{k})\delta_{ss'}\delta(\boldsymbol{k}-\boldsymbol{k}') - Z^{-1}\int \mathrm{d}^4 x\,\mathrm{d}^4 y\,\mathrm{e}^{\mathrm{i}(k'\cdot x - k\cdot y)}$$

$$\times\ \bar{u}(k', s')(\mathrm{i}\not{\partial}_x - m)\langle 0|T[\psi(x)\bar{\psi}(y)]|0\rangle(-\mathrm{i}\overleftarrow{\not{\partial}}_y - m)u(k, s). \tag{9.16}$$

Included in the definition of the T product is a change of sign for each interchange of a pair of fermion fields needed to bring the initial product into the correct time order.

By means of reduction formulae, all probability amplitudes for collision and decay processes can be expressed in terms of vacuum expectation values of time-ordered products of field operators. Except in very special cases, these expectation values can be calculated only approximately. Suitable methods of approximation can be developed by continuing to work with field operators, but a much more convenient framework for calculation is available, namely the *path integral* formalism, which I shall now develop.

9.3 Path Integrals

9.3.1 Path integrals in non-relativistic quantum mechanics

To reduce things to their simplest terms, consider first the non-relativistic theory of a single particle, moving in one dimension in a potential $V(x)$. To make the analogy with field theory as close as possible, I will take the mass of the particle to be $m = 1$. A quantity somewhat analogous to the Green functions of quantum field theory is the matrix element

$$G_{\mathrm{fi}}(t_1, t_2) = \langle x_{\mathrm{f}}, t_{\mathrm{f}}|T[\hat{x}(t_1)\hat{x}(t_2)]|x_{\mathrm{i}}, t_{\mathrm{i}}\rangle. \tag{9.17}$$

The ket $|x_{\mathrm{i}}, t_{\mathrm{i}}\rangle$ is a Heisenberg-picture vector representing that history in which the particle is at the point x_{i} at the initial time t_{i} (but may be found elsewhere at other times), so it is an eigenvector of the Heisenberg-picture operator $\hat{x}(t)$ at the instant $t = t_{\mathrm{i}}$ only. The bra $\langle x_{\mathrm{f}}, t_{\mathrm{f}}|$ is defined similarly, and t_1 and t_2 lie between t_{i} and t_{f}. To be definite, let us take the Heisenberg and Schrödinger pictures to coincide at $t = t_{\mathrm{i}}$, which means that

$$\hat{x}(t) = \exp[\mathrm{i}\hat{H}(t - t_{\mathrm{i}})]\hat{x}\exp[-\mathrm{i}\hat{H}(t - t_{\mathrm{i}})]. \tag{9.18}$$

A little thought will show that, since $\hat{x}(t_{\mathrm{f}})|x_{\mathrm{f}}, t_{\mathrm{f}}\rangle = x_{\mathrm{f}}|x_{\mathrm{f}}, t_{\mathrm{f}}\rangle$, the dependence of this eigenvector on t_{f} is given by

$$|x_{\mathrm{f}}, t_{\mathrm{f}}\rangle = \exp[\mathrm{i}\hat{H}(t_{\mathrm{f}} - t_{\mathrm{i}})]|x_{\mathrm{f}}\rangle \tag{9.19}$$

which is different from the time dependence of a Schrödinger-picture state vector such as (5.29).

The idea of a path integral, due to P A M Dirac and R P Feynman, is that the matrix element (9.17) can be expressed as an integral over all paths $x(t)$ that the particle might follow between x_{i} at time t_{i} and x_{f} at time t_{f}. An integral over

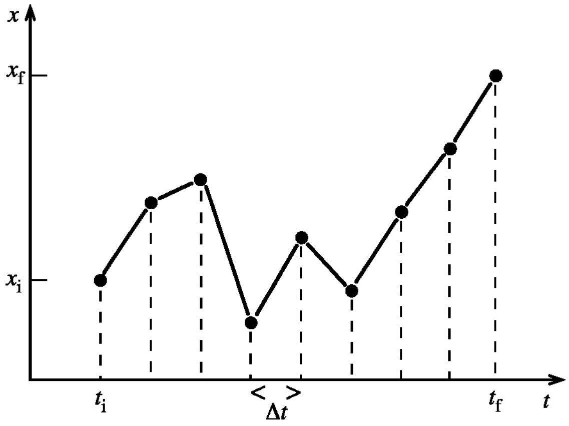


Figure 9.1. Construction of a Feynman path integral over all trajectories leading from $x_{\rm i}$ at time $t_{\rm i}$ to $x_{\rm f}$ at time $t_{\rm f}$.

paths can be defined by splitting the time interval $t_{\rm f} - t_{\rm i}$ into N segments, each of length Δt, doing an ordinary multiple integral over the $N - 1$ points $x(t_{\rm i} + n\Delta t)$ and taking the limit $N \to \infty$, as illustrated in figure 9.1.

Symbolically, this may be written as

$$\int \mathcal{D}x(t)\,(\ldots) = \lim_{N\to\infty} \int_{-\infty}^{\infty} \prod_{n=1}^{N-1} \mathrm{d}x_n\,(\ldots). \tag{9.20}$$

(A somewhat more rigorous treatment can be given in terms of probability measures over suitable classes of functions.)

To see how (9.17) can be expressed in terms of such an integral, we first translate it into the Schrödinger picture. For the case $t_2 > t_1$, we have

$$G_{\rm fi}(t_1, t_2) = \langle x_{\rm f}|\exp[-\mathrm{i}\hat{H}(t_{\rm f} - t_2)]\hat{x}\exp[-\mathrm{i}\hat{H}(t_2 - t_1)]\hat{x}\exp[-\mathrm{i}\hat{H}(t_1 - t_{\rm i})]|x_{\rm i}\rangle. \tag{9.21}$$

Now, $|x\rangle$ is an eigenvector of the Schrödinger-picture operator $\hat{x}$, so we can use the results of exercise 5.4 to write

$$\hat{I} = \int_{-\infty}^{\infty} \mathrm{d}x\,|x\rangle\langle x| \qquad \hat{x} = \int_{-\infty}^{\infty} \mathrm{d}x\,|x\rangle x\langle x| \tag{9.22}$$

where $\hat{I}$ is the identity operator. Making use of the second of these, (9.21) becomes.

$$\int_{-\infty}^{\infty} \mathrm{d}x_1 \mathrm{d}x_2\,\langle x_{\rm f}|\exp[-\mathrm{i}\hat{H}(t_{\rm f} - t_2)]|x_2\rangle x_2 \\ \times \langle x_2|\exp[-\mathrm{i}\hat{H}(t_2 - t_1)]|x_1\rangle x_1\langle x_1|\exp[-\mathrm{i}\hat{H}(t_1 - t_{\rm i})]|x_{\rm i}\rangle. \tag{9.23}$$

In the same way, we can split up each of the remaining matrix elements into a large number of short time intervals, this time using repeated insertions of $\hat{I}$:

$$\begin{aligned}&\langle x_\mathrm{f}|\exp[-\mathrm{i}\hat{H}(t_\mathrm{f}-t_\mathrm{i})]|x_\mathrm{i}\rangle\\&\qquad=\int_{-\infty}^{\infty}\prod_{n=1}^{N-1}\mathrm{d}x_n\,\langle x_\mathrm{f}|\mathrm{e}^{-\mathrm{i}\hat{H}\Delta t}|x_{N-1}\rangle\cdots\langle x_1|\mathrm{e}^{-\mathrm{i}\hat{H}\Delta t}|x_\mathrm{i}\rangle.\end{aligned}\tag{9.24}$$

We now need to evaluate each of the matrix elements on the right-hand side. The following is a rough-and-ready method that gives the right answer, but more sophisticated treatments are possible. If $\Delta t = (t_\mathrm{f} - t_\mathrm{i})/N$ is small enough, the exponential in each matrix element can be expanded as

$$\langle x_2|\mathrm{e}^{-\mathrm{i}\hat{H}\Delta t}|x_1\rangle \approx \langle x_2|\left[\hat{I}-\tfrac{1}{2}\mathrm{i}\Delta t\,\hat{p}^2-\mathrm{i}\Delta t\,V(\hat{x})\right]|x_1\rangle.\tag{9.25}$$

Taking each operator in turn, we can evaluate the matrix elements as

$$\begin{aligned}\langle x_2|\hat{I}|x_1\rangle &= \delta(x_2-x_1) = (2\pi)^{-1}\int \mathrm{d}k\ \exp[\mathrm{i}k(x_1-x_2)]\\\langle x_2|V(\hat{x})|x_1\rangle &= (2\pi)^{-1}\int \mathrm{d}k\ \exp[\mathrm{i}k(x_1-x_2)]V(x_2)\\\langle x_2|\hat{p}^2|x_1\rangle &= (2\pi)^{-1}\int \mathrm{d}k\,\mathrm{d}k'\ \exp[\mathrm{i}(kx_1-k'x_2)]\langle k'|\hat{p}^2|k\rangle\\&= (2\pi)^{-1}\int \mathrm{d}k\ \exp[\mathrm{i}k(x_1-x_2)]k^2.\end{aligned}$$

On re-exponentiating, we find

$$\langle x_2|\mathrm{e}^{-\mathrm{i}\hat{H}\Delta t}|x_1\rangle = (2\pi)^{-1}\int \mathrm{d}k\ \exp[\mathrm{i}k(x_1-x_2)-\tfrac{1}{2}\mathrm{i}\Delta t\,k^2-\mathrm{i}\Delta t\,V(x_2)]\tag{9.26}$$

up to terms of order $(\Delta t)^2$. We now shift the integration variable by $k \rightarrow k + (x_1 - x_2)/\Delta t$, after which the k integral produces just a constant:

$$\langle x_2|\mathrm{e}^{-\mathrm{i}\hat{H}\Delta t}|x_1\rangle \approx \text{constant}\times\exp\left\{\mathrm{i}\Delta t\left[\frac{1}{2}\left(\frac{x_1-x_2}{\Delta t}\right)^2-V(x_2)\right]\right\}.\tag{9.27}$$

In the limit $\Delta t \rightarrow 0$, this becomes exact, so for a longer time interval we can use (9.24) to write

$$\begin{aligned}&\langle x_\mathrm{f}|\exp[-\mathrm{i}\hat{H}(t_\mathrm{f}-t_\mathrm{i})]|x_\mathrm{i}\rangle = \text{constant}\\&\quad\times\lim_{N\to\infty}\int_{-\infty}^{\infty}\prod_{n=1}^{N-1}\mathrm{d}x_n\ \exp\left\{\mathrm{i}\Delta t\sum_{n=1}^{N}\left[\frac{1}{2}\left(\frac{x_n-x_{n-1}}{\Delta t}\right)^2-V(x_n)\right]\right\}\end{aligned}\tag{9.28}$$

where $x_0 = x_{\mathrm{i}}$ and $x_N = x_{\mathrm{f}}$. Let us now consider the points x_n to belong to a path $x(t)$, with $x_n = x(t_{\mathrm{i}} + n\Delta t)$. Then $(x_n - x_{n-1})/\Delta t = \dot{x}(t)$, and we recognize the expression in square brackets in (9.28) as the classical Lagrangian

$$L = \tfrac{1}{2}\dot{x}^2 - V(x). \tag{9.29}$$

When we apply this result to (9.23), x_1 and x_2 become $x(t_1)$ and $x(t_2)$ respectively, and we get the result

$$\langle x_{\mathrm{f}}|\hat{x}(t_2)\hat{x}(t_1)|x_{\mathrm{i}}\rangle = \int \mathcal{D}x(t)\, x(t_1)x(t_2) \exp\left(\mathrm{i}\int_{t_{\mathrm{i}}}^{t_{\mathrm{f}}} L(\dot{x}, x)\mathrm{d}t\right) \tag{9.30}$$

where the path integral is over all paths for which $x(t_{\mathrm{i}}) = x_{\mathrm{i}}$ and $x(t_{\mathrm{f}}) = x_{\mathrm{f}}$, and all the constants from k integrations have been absorbed into the definition of the symbol $\mathcal{D}x(t)$. A close inspection of steps we have gone through should reveal that (9.30) is valid only when $t_2 > t_1$. On the right-hand side, however, $x(t_1)$ and $x(t_2)$ are ordinary commuting numbers, so the order in which they are written down does not matter. Therefore, if $t_1 > t_2$, we would obtain exactly the same result for the quantity $\langle x_{\mathrm{f}}|\hat{x}(t_1)\hat{x}(t_2)|x_{\mathrm{i}}\rangle$. In other words, the path-integral we have derived actually represents the matrix element of the *time-ordered* product from which we started. Readers should not find it hard to convince themselves that the general result

$$\langle x_{\mathrm{f}}|T[\hat{x}(t_1)\cdots\hat{x}(t_n)]|x_{\mathrm{i}}\rangle \\ = \int \mathcal{D}x(t)\, x(t_1)\cdots x(t_n) \exp\left(\mathrm{i}\int_{t_{\mathrm{i}}}^{t_{\mathrm{f}}} L(\dot{x}, x)\mathrm{d}t\right) \tag{9.31}$$

can be obtained in just the same way.

9.3.2 Functional integrals in quantum field theory

Despite some slight technical complications that I shall not go into, the vacuum expectation values of time-ordered products of field operators which appear in the reduction formulae for S-matrix elements can be represented by integrals similar to (9.31). For a scalar field, we have

$$\langle 0|T[\hat{\phi}(x_1)\cdots\hat{\phi}^{\dagger}(x_n)]|0\rangle = \int \mathcal{D}\phi(x)\, \phi(x_1)\cdots\phi^*(x_n)\mathrm{e}^{\mathrm{i}S(\phi)} \tag{9.32}$$

where $S(\phi)$ is the action. The integral is over complex functions $\phi(x)$ and is often called a *functional integral* rather than a path integral. The adjoint field operator $\hat{\phi}^{\dagger}(x)$ is represented in the integral by the complex conjugate function $\phi^*(x)$, and if the field is Hermitian the integral is only over real functions.

In the case of fermions, the fields in the functional integral must be taken as Grassmann variables, to take account of the anticommuting properties of the original field operators. I give a brief discussion of the properties of Grassmann integrals in appendix A and further details may be found in specialized field theory textbooks, but few of these details will be needed for the purposes of this chapter.

It might seem that functional integrals would be extremely difficult to evaluate and so, more often than not, they are. In practice, however, it is often possible to extract the results we require by means of manipulations that avoid our having to compute a functional integral directly. As a first example, let us evaluate the Feynman propagator (9.14) for a free scalar field. It is convenient to introduce a *generating functional* for the Green functions (9.32), defined by

$$Z_0(J, J^*) = \int \mathcal{D}\phi \, \exp\left\{ \mathrm{i} \int \mathrm{d}^4x \left[\mathcal{L}_0 + J^*(x)\phi(x) + J(x)\phi^*(x) \right] \right\} \tag{9.33}$$

where $\mathcal{L}_0$ is the free-field Lagrangian density (7.7) and the definition of the measure $\mathcal{D}\phi$ is adjusted by a constant factor so that $Z_0(0, 0) = 1$. The propagator is given by

$$G_{\mathrm{F}}(x - y) = \mathrm{i} \frac{\delta}{\delta J^*(x)} \frac{\delta}{\delta J(y)} Z_0(J, J^*) \bigg|_{J=J^*=0} \tag{9.34}$$

and other Green functions can obviously be generated by further differentiations. The functional derivative $\delta/\delta J(x)$ works in much the same way as a partial derivative and is explained in detail in appendix A. The quantities $J(x)$ and $J^*(x)$, usually called *sources*, serve as a mathematical book-keeping device and have no direct physical meaning.

In this and other calculations, it is necessary to re-express spacetime integrals using integrations by parts. For simplicity, I shall usually assume that boundary conditions can be applied which ensure that surface terms vanish. Readers should be aware, however, that this cannot always be done. In particular, the nonlinear field equations, which are the Euler–Lagrange equations of interacting field theories, frequently have topologically non-trivial solutions, described in the literature as solitons, monopoles, instantons, vortices and the like (see chapter 13). When these are important, the boundary conditions must be considered more carefully. With this proviso, the exponent in (9.33) can be written in the form

$$-\mathrm{i} \int \mathrm{d}^4x \, \Phi^*(x)(\Box_x + m^2)\Phi(x) + \mathrm{i} \int \mathrm{d}^4x \, \mathrm{d}^4y \, J^*(x) g(x - y) J(y) \tag{9.35}$$

where

$$\Phi(x) = \phi(x) + \int \mathrm{d}^4y \, g(x - y) J(y) \tag{9.36}$$

and $g(x - y)$ is a Green function which satisfies the equation

$$(\Box + m^2) g(x - y) = -\delta(x - y). \tag{9.37}$$

Since the functional integral over ϕ is the limit of a product of ordinary integrals with the range $-\infty$ to ∞, we can shift the integration variable by an amount that does not depend on ϕ (say, by $-\int \mathrm{d}^4y \, g(x - y) J(y)$) without changing the value of the integral. In effect, it is equivalent to an integral over Φ, which contributes to Z_0 a factor independent of J and J^*. Thus, we have found

$$Z_0(J, J^*) = Z_0(0, 0)\exp\left[-\mathrm{i}\int \mathrm{d}^4x \int \mathrm{d}^4y\, J^*(x)g(x-y)J(y)\right]. \tag{9.38}$$

In view of the normalization $Z(0, 0) = 1$, we have succeeded in evaluating the generating functional without actually carrying out a functional integral, as long as we can find the function $g(x - y)$. It is easy to verify that $g(x - y)$ can be expressed as a Fourier transform

$$g(x-y) = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2 - m^2}. \tag{9.39}$$

This is not well defined as it stands, however, because the integrand has poles at $k^0 = \pm(\boldsymbol{k}^2 + m^2)^{1/2}$. In fact, if the k^0 integral is carried out as a contour integral, then several different solutions to (9.37) can be found by routing the contour round the poles in different ways. Equivalently, each pole can be shifted into the complex plane by a small amount $\pm\mathrm{i}\epsilon$, which is taken to zero after the integration, and different choices of the ± signs give different solutions of (9.37). Now, according to (9.34), the Feynman propagator is equal to $g(x - y)$, so we must choose that solution which agrees with the original definition (9.14). In the free field theory, this can be calculated directly using the properties of the field operators, and the correct definition is found to be (see exercise 9.3)

$$G_{\mathrm{F}}(x-y) = \lim_{\epsilon\to 0}\int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2 - m^2 + \mathrm{i}\epsilon}. \tag{9.40}$$

Our final result for the generating functional is therefore

$$Z_0(J, J^*) = \exp\left[-\mathrm{i}\int \mathrm{d}^4x \int \mathrm{d}^4y\, J^*(x)G_{\mathrm{F}}(x-y)J(y)\right]. \tag{9.41}$$

The appearance of this prescription of replacing m^2 by $m^2 - \mathrm{i}\epsilon$ may be understood as follows. The functional integral (9.33) is not really well defined, because the magnitude of the integrand is, for any value of ϕ, a complex number of unit magnitude. In effect, the $m^2 - \mathrm{i}\epsilon$ prescription adds to the exponent a term $-\epsilon\int \mathrm{d}^4x\, |\phi(x)|^2$. This provides a convergence factor, which makes the integrand decay to zero at large values of $|\phi|$.

For spin-$\frac{1}{2}$ particles, the Feynman propagator is a 4 × 4 matrix defined by

$$S_{\mathrm{F}ij}(x-y) = -\mathrm{i}\langle 0|T[\psi_i(x)\bar{\psi}_j(y)]|0\rangle. \tag{9.42}$$

It satisfies the spinor version of (9.37), namely

$$(\mathrm{i}\not{\partial} - m)S_{\mathrm{F}}(x-y) = \delta(x-y) \tag{9.43}$$

and is given by

$$\begin{aligned} S_{\mathrm{F}}(x-y) &= (\mathrm{i}\partial\!\!\!/ + m)G_{\mathrm{F}}(x-y) \\ &= \lim_{\epsilon\to 0}\int \frac{\mathrm{d}^4x}{(2\pi)^4}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\frac{(k\!\!\!/ + m)}{k^2-m^2+\mathrm{i}\epsilon}. \end{aligned} \tag{9.44}$$

9.4 Perturbation Theory

The simplest example of an interacting field theory is a scalar field theory whose Lagrangian density has the form $\mathcal{L} = \mathcal{L}_0 - V(\phi, \phi^*)$, where V is a polynomial in the fields ϕ and ϕ^*. The generating functional for its Green functions can be written as

$$\begin{aligned} Z(J, J^*) &= \int \mathcal{D}\phi \exp\left[\mathrm{i}\int \mathrm{d}^4x(\mathcal{L} + J^*\phi + J\phi^*)\right] \\ &= \int \mathcal{D}\phi \exp\left[-\mathrm{i}\int \mathrm{d}^4x\, V(\phi,\phi^*)\right]\exp\left[\mathrm{i}\int \mathrm{d}^4x(\mathcal{L}_0 + J^*\phi + J\phi^*)\right]. \end{aligned} \tag{9.45}$$

In the second form, differentiation of $\exp[\mathrm{i}\int \mathrm{d}^4x(\mathcal{L}_0 + J^*\phi + J\phi^*)]$ with respect to $J(x)$ or $J^*(x)$ multiplies it by $\mathrm{i}\phi^*(x)$ or $\mathrm{i}\phi(x)$, so we can also write this as

$$Z(J, J^*) = N\exp\left[-\mathrm{i}\int \mathrm{d}^4x\, V\left(-\mathrm{i}\frac{\delta}{\delta J^*(x)}, -\mathrm{i}\frac{\delta}{\delta J(x)}\right)\right] Z_0(J, J^*) \tag{9.46}$$

where N is a normalizing constant determined by the condition $Z(0, 0) = 1$. The most useful theory of this kind is defined by

$$V(\phi, \phi^*) = \tfrac{1}{4}\lambda(\phi^*\phi)^2 \tag{9.47}$$

where λ is a coupling constant, which determines the strength of the force acting on ϕ particles in much the same way that electric charge determines the strength of electromagnetic forces. There is no known way of computing this generating functional or any of the individual Green functions exactly. A commonly used method of approximation is *perturbation theory*, which means an expansion in powers of λ. To see how this expansion works, let us first calculate the normalization factor N in (9.46). On expanding the exponential and setting $Z_0(0, 0) = 1$, we obtain

$$\begin{aligned} &Z(0,0) \\ &\quad = N\left[1 - \frac{\mathrm{i}}{4}\lambda\int \mathrm{d}^4x\left(\frac{\delta}{\delta J(x)}\right)^2\left(\frac{\delta}{\delta J^*(x)}\right)^2 Z_0(J, J^*)\bigg|_{J=J^*=0} + \mathrm{O}(\lambda^2)\right]. \end{aligned} \tag{9.48}$$

When the expression (9.41) for Z_0 is expanded in powers of J and J^*, we see that after differentiating and setting $J = J^* = 0$, only the term containing $(\int J^* G_{\mathrm{F}} J)^2$ survives. By carrying out the functional differentiation, we find that the normalizing constant is

$$N = 1 - \tfrac{1}{2}\mathrm{i}\lambda \int \mathrm{d}^4x \, [G_{\mathrm{F}}(0)]^2 + \mathrm{O}(\lambda^2). \tag{9.49}$$

Taking this result into account, we can find a similar approximation to the propagator of the interacting theory, defined by

$$G(x-y) = -\mathrm{i}\langle 0|T[\phi(x)\phi^\dagger(y)]|0\rangle = \mathrm{i}\frac{\delta}{\delta J^*(x)}\frac{\delta}{\delta J(y)} Z(J, J^*)\bigg|_{J=J^*=0} \tag{9.50}$$

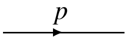
which is

$$\begin{aligned} G(x-y) = {} & G_{\mathrm{F}}(x-y) \\ & + \lambda \int \mathrm{d}^4z \, (-\mathrm{i})^3 G_{\mathrm{F}}(x-z) G_{\mathrm{F}}(z-z) G_{\mathrm{F}}(z-y) + \mathrm{O}(\lambda^2). \end{aligned} \tag{9.51}$$

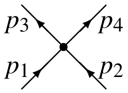
Its Fourier transform can be written, using (9.40), as

$$\begin{aligned} \widetilde{G}(p) = {} & \frac{1}{p^2 - m^2 + \mathrm{i}\epsilon} \\ & + \frac{\mathrm{i}\lambda}{(p^2 - m^2 + \mathrm{i}\epsilon)^2} \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{1}{k^2 - m^2 + \mathrm{i}\epsilon} + \mathrm{O}(\lambda^2). \end{aligned} \tag{9.52}$$

This expression, and those arising in the perturbation series for all other Green functions, are conveniently represented by *Feynman diagrams*. The diagrams corresponding to (9.52) are shown in figure 9.2, and are constructed according to the following rules:

(i) 

stands for $\dfrac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}$

(ii) 

stands for $-\mathrm{i}\lambda$, together with the condition $p_1 + p_2 = p_3 + p_4$ for momentum conservation

(iii) All internal momenta, such as k in figure 9.2, whose values are not fixed by momentum conservation are integrated over.

(iv) Each diagram has a combinatorial factor arising from the expansions of exponentials and the chain rule for differentiation. Many field theory textbooks supply rules for calculating this factor, but in my experience it is best obtained from first principles.

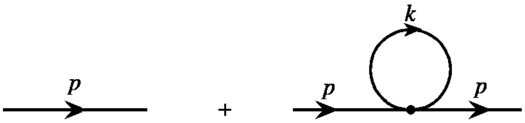


Figure 9.2. Diagrammatic representation of equation (9.52).

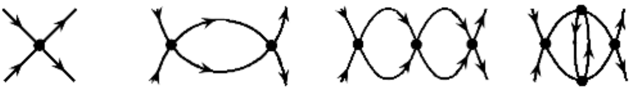
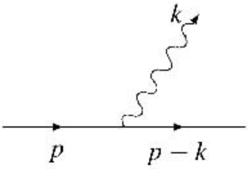


Figure 9.3. Examples of Feynman diagrams which contribute to the elastic scattering amplitude for two spin-0 particles.

At a given order in λ, there are fixed numbers of vertices and unperturbed propagators available, and there is a contribution to the Green function from each diagram that can be formed from these elements. For example, figure 9.3 shows some of the diagrams which contribute to the S-matrix element for two-particle elastic scattering. Each diagram has four external propagators, one for each of the two incoming and two outgoing particles. The S-matrix element itself is similar to (9.15), but with a factor $(k_i^2 - m^2)$ for each particle multiplying the Green function. Evidently, these are just cancelled by the external propagators in figure 9.3, leaving a non-zero result.

The Feynman rules for theories containing fermions differ in two respects from those given above. One is that each propagator line represents the matrix (9.44). Most often, only two fermion lines meet at any given vertex. For example, a term $e\bar{\psi}\not{A}\psi$ in the action (8.17) gives rise to a vertex of the form



where the wavy line denotes the photon propagator, to be discussed in the next section. As far as the fermion is concerned, this vertex, together with the propagators, corresponds to the matrix product

$$-\mathrm{i}eS_{\mathrm{F}}(p)\gamma^{\mu}S_{\mathrm{F}}(p-k) \tag{9.53}$$

whose index μ will be contracted with a corresponding index belonging to the photon propagator. Each internal fermion propagator will be multiplied by a matrix on either side. An external propagator will be multiplied by a matrix on one side (where it meets a vertex) leaving one free Dirac index. This free index is the one belonging to a field operator in the original matrix element, such as (9.42), and will eventually be contracted with a Dirac operator and a wavefunction, as in the reduction formula (9.16). The second difference is the appearance of some power of -1 in the combinatorial factor. These signs arise from the anticommutation properties of the Grassmann variables in the functional integral. Every closed loop of fermion propagators gives a factor of -1 and extra minus signs come from the ordering of field operators in a time-ordered product. Once again, I must ask readers who wish to become proficient in these calculations to consult a specialized text for details of the technicalities.

Feynman diagrams such as those in figures 9.2 and 9.3 are often thought of as representing actual physical processes. For example, the first diagram of figure 9.3 might be thought of as an immediate transition from the initial two-particle state to the final two-particle state, while the higher-order diagrams represent indirect transitions via the intermediate states in which particles corresponding to the internal propagators are created and subsequently annihilated. The net effect of each of these processes is the same, in the sense that they each involve the same initial and final states. The overall probability amplitude is the sum of the amplitudes for all possible ways in which this net transition can occur. A particle whose transitory existence is represented by an internal propagator differs from a real, observable particle, because its 4-momentum does not have to satisfy the mass-shell constraint $k^2 = m^2$. For this reason, the intermediate particles are called *virtual particles*. The idea of virtual particles provides a pictorial language that is often useful for discussing the mathematics of perturbation theory. Clearly, however, this language is closely tied to our use of an expansion in powers of λ or some other coupling constant; the notion of virtual particles is meaningful, at best, only when perturbation theory gives an accurate approximation to the observable quantities we are attempting to calculate.

9.5 Quantization of Gauge Fields

We saw in §7.6 that gauge fields, such as the electromagnetic 4-vector potential A_μ, can be treated as field operators whose associated particles are vector bosons, such as the photon. However, there are problems in the quantum-mechanical treatment that do not arise for scalar or spinor fields and which are most conveniently overcome by the use of path integrals. Symptomatic of these problems is the fact that, although A_μ has four components, photons exist, as we have seen, in only two independent helicity states. Therefore, two of the four field degrees of freedom are in some way redundant, being unobservable gauge degrees of freedom.

Mathematically, this can be seen as follows. In the absence of charged particles, the action of electromagnetism is the first term of (8.17). With $F_{\mu\nu}$

given by the antisymmetric expression $\partial_\mu A_\nu - \partial_\nu A_\mu$, this action is independent of $\partial_0 A_0$ and therefore, as indicated in exercise 3.6, the canonical momentum Π^0 conjugate to A_0 is identically zero. Thus, there are at most three independent momenta

$$\Pi^i = \frac{\delta S}{\delta(\partial_0 A_i)} = F^{i0} = E^i. \tag{9.54}$$

Since there are at most three independent momenta, there can also be at most three independent field variables. To reduce the matter to its simplest terms, let us regard A_0 as the redundant component. The four Euler–Lagrange equations are Maxwell's equations $\partial_\mu F^{\mu\nu} = 0$. The one obtained by varying A_0 cannot be regarded as an equation of motion on the same footing as the others, because A_0 is not a *bona fide* dynamical variable, but must rather be regarded as a further constraint on the remaining field components. (Readers familiar with such matters will realize that A_0 is playing the role of a Lagrange multiplier.) The offending Maxwell equation is Gauss' law (3.44) which, given (9.54), may be written as

$$\nabla \cdot \boldsymbol{E} = \partial_i \Pi^i = 0. \tag{9.55}$$

Clearly, this equation has no time derivatives and is not a genuine equation of motion. It is a relation between the three momenta, which implies that only two of these momenta are really independent. We conclude that there are really only two genuine field variables and two conjugate momenta, corresponding to the two observed polarization states of the photon.

For scattering processes that involve photons in the initial or final state, reduction formulae similar to (9.13) or (9.16) can be derived in which the contribution from a photon is

$$\mathrm{i}Z^{-1/2} \int \mathrm{d}^4x\, \mathrm{e}^{\pm \mathrm{i}k\cdot x} \langle 0|T[\cdots \epsilon(k)\cdot j_\mathrm{e}(x) \cdots]|0\rangle. \tag{9.56}$$

The current density $j_\mathrm{e}^\mu(x)$ is given in terms of field operators for whatever charged particles are present (for example, $j_\mathrm{e}^\mu = q\bar{\psi}\gamma^\mu\psi$ for spin-$\frac{1}{2}$ particles of charge q) and $\epsilon^\mu(k)$ is the polarization vector introduced in §7.6.1. This could have been written in a form more similar to (9.13). Indeed, with charged particles present, Maxwell's equations are

$$\Box A^\mu - \partial^\mu(\partial_\nu A^\nu) = j_\mathrm{e}^\mu \tag{9.57}$$

and we can simply substitute the expression on the left-hand side for j_e^μ. The advantage of (9.56) is that it avoids certain ambiguities concerning the definition of time-ordered products of gauge fields, as well as the question of whether the constraint $\partial_\nu A^\nu = 0$ is to be imposed and, if so, how. In terms of Feynman diagrams, j_e^μ introduces into any diagram the vertex (9.53) without the external photon propagator, so in effect we have simply cancelled out this propagator before evaluating the Green function rather than afterwards.

The reduction formula (9.56) serves to make contact with observable physical processes in a way that temporarily avoids the difficulties associated with quantizing the gauge field, but these difficulties can no longer be avoided when we come to calculate the vacuum expectation value itself, because we expect Feynman diagrams to contain internal photon or other gauge-field propagators as well as external ones. In the case of a scalar field, whose quantum-mechanical properties are straightforward, the path-integral representation (9.32) could be deduced from the canonical formalism of field operators. With enough care, the same thing can be done for a gauge field. However, it is possible to adopt an alternative point of view, regarding a path integral such as (9.32) as *defining* a quantum theory, given that we have an action S which specifies the corresponding classical theory. This *path integral quantization scheme* is an alternative to the canonical scheme of §5.4, upon which our theory up to this point has rested. If we adopt this point of view then, at first sight, it appears that we simply have to base our calculations on an appropriate generating functional

$$Z(\text{sources}) = \int \mathcal{D}(\text{fields})\mathrm{e}^{\mathrm{i}S+\text{source terms}}. \tag{9.58}$$

The functional integral is over all the fields in the theory, and the source terms are similar to those in (9.33), namely

$$\mathrm{i}\int \mathrm{d}^4x\,[J^*\phi + J\phi^* + J^\mu A_\mu + \bar{\eta}\psi + \bar{\psi}\eta + \ldots] \tag{9.59}$$

with one terms for each field. The sources J, J^μ, η, etc are the arguments of Z, and the action is an expression such as (8.17) or (8.41), perhaps with the addition of scalar fields, depending on the particular theory considered.

We should, of course, be suspicious of this procedure if it enabled us to ignore entirely the problems associated with redundant gauge degrees of freedom. In fact, these problems reappear in the following way. Since the action S is gauge invariant, the integrand in (9.58) is independent of the gauge degrees of freedom when we set the sources to zero, and the functional integrals over these degrees of freedom lead to a meaningless infinity. It is, in fact, impossible to do perturbation theory with (9.58) as it stands, because we cannot find propagators for the gauge fields. In the case of electromagnetism, if we follow the same steps as for the scalar field, we find that the propagator, denoted by $D_{\mathrm{F}\mu\nu}$, should satisfy an equation similar to (9.37), but with the Klein–Gordon operator replaced by the Maxwell operator:

$$\Box D_{\mathrm{F}\mu\nu}(x-y) - \partial_\mu\partial^\lambda D_{\mathrm{F}\lambda\nu}(x-y) = \eta_{\mu\nu}\delta(x-y). \tag{9.60}$$

This equation has no solution.

A way round this difficulty was found by L D Fadeev and V N Popov. Their argument is slightly complicated, and here I shall just state the result, but a related calculation is described in detail in §15.3.2. It is possible to modify the action by adding two terms to the Lagrangian density of the gauge fields:

$$\mathcal{L}_{\mathrm{FP}} = -\tfrac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - \tfrac{1}{2}\xi^{-1}f(A) + \bar{b}\Delta(A)c. \tag{9.61}$$

The function $f(A)$ is a function of the gauge fields, whose purpose is to remove the gauge invariance of the original action, thereby allowing a propagator to be constructed. We are allowed a considerable freedom in choosing this function, although only a limited number of choices are convenient in practice. The new fields $\bar{b}$ and c, which are to be integrated over in the generating functional, correspond to fictitious particles, usually called *ghosts*. Although these are spin-0 particles, the mathematics requires their fields to be Grassmann variables, so they are fermions, contradicting the spin-statistics theorem which applies to all physical particles. The quantity $\Delta(A)$ is a differential operator, whose exact form depends on our choice of $f(A)$. In the case of electromagnetism (where, of course, the index a does not appear), a convenient choice of $f(A)$ is

$$f(A) = (\partial_\mu A^\mu)^2. \tag{9.62}$$

With this choice, Δ turns out to be independent of A_μ. In this case, the ghosts do not interact with other particles and can be ignored. By modifying the action in this way, we naturally modify the Green functions as well. In particular, they now depend on the arbitrary parameter ξ. As a consequence of the original gauge invariance, however, it can be shown that S-matrix elements and other physically measurable gauge-invariant quantities are unaffected by the modification and are independent of ξ. I shall give an example of this is due course. The $f(A)$ term in (9.61) is often referred to as a *gauge-fixing* term. This is somewhat misleading, as it suggests that a constraint has been applied to eliminate the redundant gauge degrees of freedom. What really happens is that these degrees of freedom, together, in general, with the ghosts conspire to have no net effect on physical quantities.

When (9.62) is used for $f(A)$, readers may readily verify that the equation for the propagator becomes

$$\Box D_{\mathrm{F}\mu\nu}(x-y) - (1-\xi^{-1})\partial_\mu\partial^\lambda D_{\mathrm{F}\lambda\nu}(x-y) = \eta_{\mu\nu}\delta(x-y) \tag{9.63}$$

and that its solution is

$$D_{\mathrm{F}\mu\nu}(x-y) = -\int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2+\mathrm{i}\epsilon}\left(\eta_{\mu\nu} + (\xi-1)\frac{k_\mu k_\nu}{k^2}\right). \tag{9.64}$$

If we include in the Lagrangian density a term $\frac{1}{2}m^2A_\mu A^\mu$, we get a theory of massive vector bosons, with the propagator

$$D_{\mathrm{F}\mu\nu}(x-y) = -\int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2-m^2+\mathrm{i}\epsilon}\left(\eta_{\mu\nu} + (\xi-1)\frac{k_\mu k_\nu}{k^2-\xi m^2}\right). \tag{9.65}$$

As it stands, such a theory is not gauge invariant, so we are not really entitled to use the extra Fadeev–Popov terms. Unlike (9.64), the propagator (9.65) has a finite limit when we remove the gauge-fixing term by taking ξ to infinity:

$$D_{\mathrm{F}\mu\nu}(x-y) = -\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2-m^2+\mathrm{i}\epsilon}\left(\eta_{\mu\nu}-\frac{k_\mu k_\nu}{m^2}\right). \tag{9.66}$$

At the level of free particles, this non-gauge-invariant theory makes good sense. As we saw in §3.7, the massive spin-1 particles have three spin polarization states and the one redundant degree of freedom is removed automatically by the constraint $\partial_\mu A^\mu = 0$, which is implicit in the equation of motion. In interacting theories, however, massive vector bosons are troublesome, as we shall shortly discover.

9.6 Renormalization

Earlier on, we derived an expression (9.52) for the first-order correction to the scalar propagator in the theory with interactions given by (9.47). This correction and further corrections at higher orders of perturbation theory are properly thought of as a *self-energy*, or as a correction to the mass of the particle brought about by the interactions. Thus, the parameter m that appears in the Lagrangian density is not the true mass of the particle. It is usually called the *bare mass*, and I shall denote it henceforth by m_0. The pole of the complete propagator must appear at the true mass shell, $p^2 = m^2$, and the 'in' and 'out' states should be defined in terms of the true mass m which therefore still appears in the reduction formulae. As we shall see below, the integral in (9.52) is purely imaginary and I shall denote it by $-\mathrm{i}\Sigma(m_0)$. Then (9.52) can be written as

$$\begin{aligned}\widetilde{G}(p) &= \frac{1}{p^2-m_0^2+\mathrm{i}\epsilon}\left(1-\lambda\frac{\Sigma}{p^2-m_0^2+\mathrm{i}\epsilon}\right)^{-1} + \mathrm{O}(\lambda^2)\\ &= \frac{1}{p^2-m_0^2-\lambda\Sigma+\mathrm{i}\epsilon} + \mathrm{O}(\lambda^2).\end{aligned} \tag{9.67}$$

This is more than a merely *ad hoc* rearrangement. Amongst the whole set of Feynman diagrams that contribute to the propagator, there is the infinite sum of diagrams shown in figure 9.4, which is easily shown to be a geometric series. Thus, the true mass is given by

$$m^2 = m_0^2 + \lambda\Sigma(m_0) + \mathrm{O}(\lambda^2). \tag{9.68}$$

This relation is said to represent *mass renormalization*.

There are two more ways in which the Lagrangian of an interacting field theory reflects only indirectly the physical phenomena that the theory describes. First of all, when we include only the lowest-order corrections to the propagator

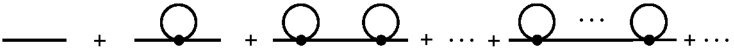


Figure 9.4. The Feynman diagrams whose sum forms the geometric series (9.67).

as in (9.67), its residue at $p^2 = m^2$ (that is, the quantity $\lim_{p^2\to m^2}(p^2-m^2)\widetilde{G}(p)$) is still equal to 1. It turns out, though, that this residue is no longer equal to 1 when higher-order corrections are also included. This means that, when acting on the vacuum state, the field operators of the interacting theory create single-particle states whose normalization is different from those of the non-interacting theory. In order to have a clear physical interpretation of our calculated scattering amplitudes, we demand that the 'in' and 'out' states should have the standard normalization of the non-interacting theory. To this end, we define the *wavefunction renormalization constant* Z, which appears in the reduction formulae, by the requirement

$$\lim_{p^2\to m^2} Z^{-1}(p^2 - m^2)\widetilde{G}(p) = 1. \tag{9.69}$$

For reasons that will shortly become apparent, it is convenient to define a renormalized field

$$\phi_{\mathrm{R}}(x) = Z^{-1/2}\phi(x) \tag{9.70}$$

and renormalized Green functions

$$G_{\mathrm{R}}^{(n)}(x_1, \ldots, x_n) = \langle 0|T[\phi_{\mathrm{R}}(x_1)\cdots\phi_{\mathrm{R}}^{\dagger}(x_n)]|0\rangle_{\mathrm{c}} \tag{9.71}$$

which take into account the adjusted normalization. (Note that the '2-point' function $G_{\mathrm{R}}^{(2)}$ also differs by a factor of $-\mathrm{i}$ from the Feynman propagator as defined in (9.14).) The subscript c here denotes the *connected* Green functions, which are obtained by ignoring all Feynman diagrams that consist of two or more disconnected parts. For example, the complete 4-point Green function (the vacuum expectation value involving four fields) contains, amongst many others, the diagrams shown in figure 9.5, but only diagrams (*a*) and (*c*) are connected. The disconnected diagrams are associated with particles that continue from the initial state to the final state without colliding, while the connected diagrams refer to particles that actually collide, and are therefore of greater interest. The complete Green functions, should we ever want them, can be expressed in terms of connected ones. In fact, it can be shown that the generating functional (9.45) is given in terms of these connected Green functions by

$$\ln Z(J, J^*) = \sum_{n=1}^{\infty} \frac{(\mathrm{i}Z^{1/2})^n}{n!} \int \mathrm{d}^4x_1 \cdots \mathrm{d}^4x_n J^*(x_1)\cdots J(x_n) G_{\mathrm{R}}^{(n)}(x_1, \ldots, x_n) \tag{9.72}$$

while $Z(J, J^*)$ itself has a similar expansion involving the complete functions.

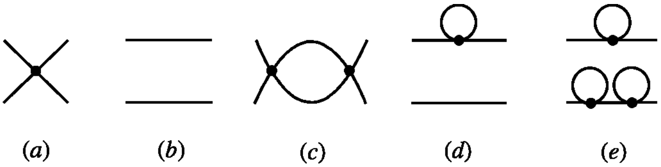


Figure 9.5. Some Feynman diagrams which contribute to the four-point Green function. Only (a) and (c) are connected diagrams.

Finally, we must recognize that the coupling constant appearing in the action, which I shall now denote by λ_0, is not a physically measurable quantity. If, for example, we measure the scattering cross-section for 2 particle → 2 particle scattering, then the measured quantity includes contributions from every Feynman diagram in $G_{\rm R}^{(4)}$; we cannot single out the contribution from diagram (a) of figure 9.5, which is simply proportional to λ_0. In order to compare the results of our calculations with experimental data, we must exchange λ_0 for a *renormalized coupling constant* λ which *is* measurable. There is considerable latitude in how we actually do this. A suitable definition might be

$$\lambda = \left[\prod_{i=1}^{4}(p_i^2 - m^2)\right] G_{\rm R}^{(4)}(p_1, \ldots, p_4)\bigg|_{p_i = p_i(\mu)} \tag{9.73}$$

where $p_i(\mu)$ are a chosen set of momentum values. These values must be specified by a parameter μ having the dimensions of momentum or equivalently, in natural units, of mass. A measurement of the cross-section for particles which have these particular momenta serves to establish the value of λ chosen by nature, and the testable content of our theory then consists in the values it predicts for the same cross-section at other momenta and for the cross-sections for other scattering processes. If we are to continue using perturbation theory, the relation between λ and λ_0 must be of the form

$$\lambda = \lambda_0 + \mathrm{O}(\lambda_0^2) \tag{9.74}$$

so that a power series in λ_0 can be re-expressed as a series in λ. The exact physical meaning of λ depends, of course, on the method used to define it and, in particular, on the chosen value of μ

The preceding remarks show, I hope, that renormalization is a natural and essential part of the physical interpretation of a quantum field theory. There is, however, a more sinister aspect to renormalization, which must now be revealed. Let us evaluate the self energy

$$\Sigma(m_0) = \mathrm{i}\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{1}{k^2 - m_0^2 + \mathrm{i}\epsilon}. \tag{9.75}$$

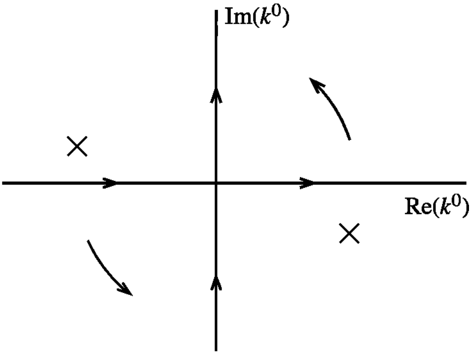


Figure 9.6. Wick rotation of the integration contour in the complex k^0 plane. Crosses mark the poles of the Feynman propagator, which do not impede the anticlockwise rotation of the contour.

If the k^0 integral is done as a contour integral, the poles in the propagator appear as in figure 9.6. The contour of integration can be rotated, avoiding these poles, to run along the imaginary axis, in effect replacing k^0 by $\mathrm{i}k^4$. The result of this process, known as a *Wick rotation*, is an integral in a four-dimensional Euclidean space, with momentum components $(k^1, \ldots, k^4)$. In this integral, the integrand depends only on the magnitude of the momentum, so in polar coordinates the angular integrations give just a constant factor. We get

$$\Sigma(m_0) = \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{1}{k^2 + m_0^2} = \frac{1}{8\pi^2} \int_0^\infty \frac{k^3 \mathrm{d}k}{k^2 + m_0^2} \tag{9.76}$$

where now $k^2 = \sum_{i=1}^4 (k^i)^2$. When k is large, the integral behaves as k^2, so it diverges quadratically at its upper limit: it is infinite! In practice, this does not matter. When we express the propagator (9.67) in terms of the true mass, it is equal to $(p^2 - m^2 + \mathrm{i}\epsilon)^{-1}$ plus higher-order corrections, and Σ does not appear in our final answer for any physical quantity. On the other hand, many other infinite integrals can be expected to occur. While these are embarrassing, we can still obtain sensible, finite results for measurable quantities provided that all infinite integrals disappear after renormalization. In quantum electrodynamics, our embarrassment is somewhat alleviated by the fact that the renormalized theory yields predictions that agree with experiment to some 10 significant figures. What we require is that the renormalized Green functions should have well-defined, finite values when they are expressed in terms of true particle masses and renormalized coupling constants. If this is true for a particular field theory, the theory is said to be *renormalizable*. It would seem that only renormalizable theories are suitable as models of physical reality, but whether this is really true is not quite clear. We are, after all, only able to do approximate calculations, and it could be that infinite answers obtained from a non-renormalizable theory are due to inadequate methods of approximation rather than to the theory itself. It practice, ways can often be found of making approximate use of a non-renormalizable theory, on the understanding that it represents only part of some more comprehensive theory.

The task of finding out whether a given field theory is renormalizable or not is a lengthy and highly technical one, and I shall do no more than state some of the essential results.

(a) A simple, though not infallible, criterion for renormalizability is provided by *dimensional analysis*. Since we are using natural units ($\hbar = c = 1$), there is only one independent unit, which I shall take to be a mass. Thus, the dimension of any quantity can be expressed as $(\text{mass})^D$. A momentum has $D = 1$. Since the two terms in a differential operator such as $(\Box + m^2)$ must have the same dimensions, ∂_μ has $D = 1$ and, correspondingly, the spacetime volume element has $D = -4$. The action S appears in a functional integral as the argument of an exponential and must therefore be dimensionless ($D = 0$), which means that $D = 4$ for a Lagrangian density. For a scalar field, whose Lagrangian density includes (7.7), this implies that $D = 1$. Similar arguments show that a gauge field also has $D = 1$, while a spinor field has $D = \frac{3}{2}$. Knowing this, it is a simple matter to determine the dimension of any coupling constant that appears in the action.

Now, the power of k with which the integral (9.76) diverges, namely 2, is equal, for fairly obvious reasons, to the dimension of the integral. Suppose, more generally, that a coupling constant λ has dimension D_λ and a Green function G has dimension D_G. We evaluate the Green function as a power series

$$G = G_0 + \lambda G_1 + \lambda^2 G_2 + \dots . \tag{9.77}$$

Each coefficient G_n is a multiple momentum integral of dimension $D_G - nD_\lambda$, which may be expected to diverge with this power. Assume that we have enough freedom, using mass, coupling constant and wavefunction renormalization, to eliminate all infinities at order $n = 1$. If D_λ is negative, the infinities become more severe at higher orders, and we might expect to reach a point where we no longer have enough freedom to eliminate them. On the other hand, if D_λ is zero or positive, then things get no worse at higher orders. A more detailed argument along these lines shows that, indeed, the theory is likely to be renormalizable if $D_\lambda \geq 0$. In fact, if D_λ is positive, then the infinities may cease altogether after some order, and the theory is said to be *super-renormalizable*. Consideration of (8.17), (8.41) and (9.47) reveals that the coupling constants e, g and λ in the theories we have thought about up to now are all dimensionless and, other things being equal, these theories should be renormalizable. One reason for restricting the actions to contain only the terms we have considered is that other possible terms would involve coupling constants of negative dimension and destroy renormalizability.

(b) When a theory possesses *symmetries* such as gauge invariance, these restrict the terms which may appear in the action, and therefore also restrict the number of independent parameters and the number of renormalizations that can be used to eliminate divergences. However, the same symmetries also restrict the ways in which infinite integrals can appear. Generally speaking, to construct a renormalizable theory, it is necessary to include in the action all possible terms that are allowed by symmetries and do not involve coupling constants of negative dimension.

(c) The dimensional criterion works for scalar field theory because the propagator $(k^2 - m^2 + \mathrm{i}\epsilon)^{-1}$ behaves for large k like k^D, where D, equal to -2, is the dimension of the propagator. The same is true of the momentum-space propagators for spin-$\frac{1}{2}$ fermions, (9.44), and photons, (9.64). For massive spin-1 particles, however, the term $k_\mu k_\nu / m^2$ in (9.66) leads to more severe divergences than are allowed for by dimensional analysis. As a result, interacting theories of massive spin-1 particles are found to be non-renormalizable, even when the dimensional criterion is satisfied. The propagator (9.65) does not lead to this problem, because of the extra power of k^2 in the denominator of the expression $k_\mu k_\nu / (k^2 - \xi m^2)$. However, the gauge-fixing term that allows us to use a propagator of this kind can be introduced only in a gauge-invariant theory. Therefore, a renormalizable theory of massive spin-1 particles must be gauge invariant. As we saw in chapter 8, special measures are necessary to achieve this.

(d) In some gauge theories which have dimensionless couplings and might be expected to be renormalizable, there occur certain ‘anomalous’ Feynman integrals whose divergences cannot be renormalized away. How and why these *anomalies* occur is the subject of a large and technical literature, the details of which I cannot pursue here. The root cause is a subtle breakdown of gauge invariance in functional integrals. Even when a fully gauge-invariant action is used, the integration measure $\mathcal{D}$(fields) in (9.58) may fail to be gauge invariant. For this reason, a field theory that is gauge invariant at the classical level may cease to be so upon quantization. Several different kinds of anomalies have been identified. The *chiral* anomalies that afflict gauge theories arise in Feynman diagrams from closed fermion loops and can be traced to gauge non-invariance of the fermionic path integral. The only way to remove them is to arrange for anomalies from several different fermion species to cancel amongst themselves. Indeed, the standard theory of weak and electromagnetic interactions (to be discussed in chapter 12) is potentially anomalous, and the sets of particle species, called *families* or *generations* of quarks and leptons, which are required for the cancellation of anomalies are exactly those whose existence is inferred from experiment.

9.7 Quantum Electrodynamics

Quantum electrodynamics, or QED for short, is the field theory that describes the behaviour of charged particles with only electromagnetic interactions. It is, of course, most useful when the effects of other interactions are negligible, and this is most nearly true when we study the properties of the charged leptons—electrons and muons. (There are also the tau particles, but these are short-lived particles produced only in high-energy collisions and their properties cannot be determined with the same accuracy.) The electrodynamics of electrons and muons is the most accurate theory in existence, if accuracy is measured by the agreement between theoretical calculations and experimental data. I shall illustrate the application of interacting field theories by discussing some well-known consequences of QED, namely the Coulomb potential, the Lamb shift of spectral lines in simple atoms, and the magnetic dipole moments of charged particles. Although the formalism has been developed with a view to interpreting scattering experiments, none of the quantities of interest here is conveniently described in these terms. Moreover, the detailed calculations involve much complicated algebra, though they are quite straightforward in principle. I shall therefore use somewhat qualitative arguments to identify the quantities that need to be calculated and omit detailed algebra when it does not illuminate questions of principle.

9.7.1 The Coulomb potential

From the point of view of perturbation theory, the interactions between charged particles come about through the exchange of virtual photons. A few of the diagrams that contribute to the scattering of two particles are shown in figure 9.7. To see how this description is related to the more elementary idea of a potential energy, let us first go back to chapter 6, where we wrote down in equation (6.21) the potential energy operator for particles interacting through a potential $V(\boldsymbol{x}, \boldsymbol{x}')$. I am going to show that all reference to photons can be eliminated from QED, leaving a theory of charged particles alone. In this version of the theory, we can, under suitable circumstances, obtain a potential energy operator of the form (6.21), which involves the familiar Coulomb potential.

For a single species of charged particle, and with the gauge-fixing function introduced in (9.62), the Lagrangian density for QED may be written as

$$\mathcal{L}_{\mathrm{QED}} = \tfrac{1}{2}A_\mu \left[\eta^{\mu\nu}\Box - (\xi^{-1} - 1)\partial^\mu\partial^\nu\right] A_\nu - j_{\mathrm{e}}^\mu A_\mu + \bar{\psi}(\mathrm{i}\not{\partial} - m)\psi \qquad (9.78)$$

where, for particles of charge q, the electromagnetic current is

$$j_{\mathrm{e}}^\mu = q\bar{\psi}\gamma^\mu\psi. \qquad (9.79)$$

The idea now is to carry out the functional integral over A_μ, leaving an effective action for ψ alone:

$$\exp[\mathrm{i}S_{\mathrm{eff}}(\psi)] = \int \mathcal{D}A \exp[\mathrm{i}S(\psi, A)]. \qquad (9.80)$$

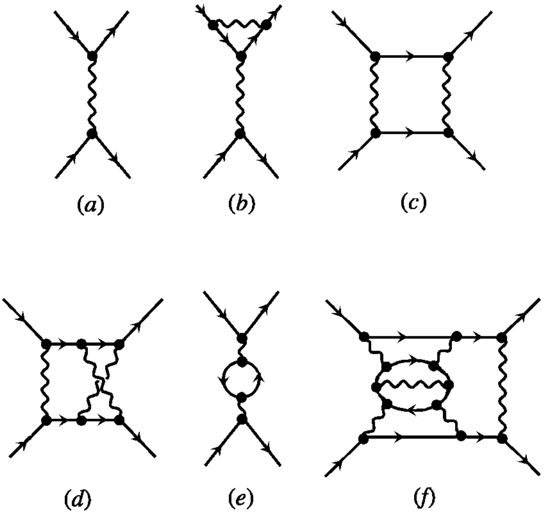


Figure 9.7. Some diagrams which contribute to the elastic scattering amplitude for two electrons. Diagrams (a) and (e) are the first two of a geometric series analogous to figure 9.4.

This is easy to do because, as far as the A integral is concerned, the current density can be considered as a source, similar to that in (9.33) or (9.59). In the same way that we derived (9.38), but using the photon propagator, we obtain

$$S_{\text{eff}} = \int \mathrm{d}^4x \, \bar{\psi}(x)(\mathrm{i}\partial\!\!\!/ - m)\psi(x) + \tfrac{1}{2}\int \mathrm{d}^4x\mathrm{d}^4y \, j_{\mathrm{e}}^{\mu}(x)D_{\mathrm{F}\mu\nu}(x-y)j_{\mathrm{e}}^{\nu}(y). \quad (9.81)$$

Obviously, we would like to identify the last term as

$$-\tfrac{1}{2}\int \mathrm{d}t \int \mathrm{d}^3x\mathrm{d}^3y \, \rho(\boldsymbol{x},t)V(\boldsymbol{x}-\boldsymbol{y})\rho(\boldsymbol{y},t) \quad (9.82)$$

where $\rho = \psi^{\dagger}\psi$ is the particle density.

The idea of a potential energy $V(\boldsymbol{x},\boldsymbol{y})$ between two particles located at $\boldsymbol{x}$ and $\boldsymbol{y}$ is really a classical one. To extract a comparable notion from the quantum-mechanical action (9.81), I shall imagine the current density (9.79) to represent an actual distribution of real charged particles although, in reality, it stands for a quantum-mechanical operator and appears only in intermediate stages of a calculation of, say, a scattering cross-section. Readers may like to consider for themselves how this step might be justified more rigorously. We can verify immediately that the effective action (9.81) is independent of the arbitrary gauge-fixing parameter ξ. This follows from the conservation of electric charge, expressed by the equation of continuity $\partial_{\mu}j_{\mathrm{e}}^{\mu} = 0$. Thus, if we insert the photon propagator (9.64) into (9.81), then $(\xi - 1)$ is multiplied by two integrals of the form

$$\int \mathrm{d}^4y\, \mathrm{e}^{\mathrm{i}k\cdot y} k_\nu j_\mathrm{e}^\nu(y) = \int \mathrm{d}^4y\, (-\mathrm{i}\partial_\nu \mathrm{e}^{\mathrm{i}k\cdot y}) j_\mathrm{e}^\nu(y) = \mathrm{i}\int \mathrm{d}^4y\, \mathrm{e}^{\mathrm{i}k\cdot y} \partial_\nu j_\mathrm{e}^\nu(y) = 0 \tag{9.83}$$

and therefore has no effect. (The second step of this calculation requires an integration by parts, with the usual assumption that $j_\mathrm{e}^\nu(y) \to 0$ for $y^\mu \to \pm\infty$.) We now obtain the standard Coulomb potential by considering a *static* distribution of charged particles, for which $j_\mathrm{e}^\mu = q(\rho, \mathbf{0})$ and the particle density ρ is independent of time. For the second term of (9.81) we then get

$$\begin{aligned}\int \mathrm{d}^4x\mathrm{d}^4y\, j_\mathrm{e}^\mu(x) D_{\mathrm{F}\mu\nu}(x-y) j_\mathrm{e}^\nu(y) &= q^2 \int \mathrm{d}^4x\mathrm{d}^4y\, \rho(\boldsymbol{x}) D_{\mathrm{F}00}(x-y)\rho(\boldsymbol{y}) \\ &= -\int \mathrm{d}t \int \mathrm{d}^3x\mathrm{d}^3y\, \rho(\boldsymbol{x}) V(\boldsymbol{x}-\boldsymbol{y})\rho(\boldsymbol{y})\end{aligned}$$

where I have written t and t' for x^0 and y^0, and the potential is

$$\begin{aligned}V(\boldsymbol{x}-\boldsymbol{y}) &= -q^2 \lim_{\epsilon\to 0} \int \mathrm{d}t' \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{\mathrm{e}^{-\mathrm{i}k_0(t-t')}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{k_0^2 - \boldsymbol{k}^2 + \mathrm{i}\epsilon} \\ &= q^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})}}{\boldsymbol{k}^2} \\ &= \frac{q^2}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}\end{aligned} \tag{9.84}$$

which is the Coulomb potential. In this calculation, I have used the fact that $\int \mathrm{d}t' \mathrm{e}^{-\mathrm{i}k_0(t-t')} = 2\pi\delta(k_0)$, and I leave it as an exercise for readers to verify the result of the final integral. If the charge distribution is not static, then the interaction cannot be described just by an electric potential. There will, for example, be corrections for the magnetic force between particles in relative motion. In the case of a force mediated by exchange of massive particles, say of mass M, we should expect the potential to be of the *Yukawa* form

$$V(\boldsymbol{r}) = q^2 \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}}{\boldsymbol{k}^2 + M^2} = \frac{q^2 \mathrm{e}^{-M|\boldsymbol{r}|}}{4\pi|\boldsymbol{r}|}. \tag{9.85}$$

The *range* of such a force, as measured by the exponential decay, is a distance equal to $1/M$ in natural units or to $\hbar/Mc$ in laboratory units. For example, this potential was suggested by Yukawa as a means of describing the strong force that binds nucleons together to form a nucleus. Assuming that the exchanged particles are pions, with masses given by $Mc^2 \approx 135\,\mathrm{MeV}$, we calculate a range of about 1.5×10^{-15} m, which is indeed typical of the separation of nucleons in a nucleus. We shall see in chapter 12, however, that the modern view of strong forces is rather more complicated than this.

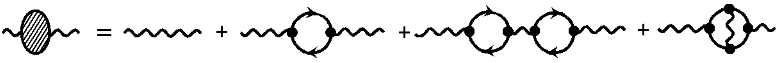


Figure 9.8. Some diagrams which contribute to the complete photon propagator.

9.7.2 Vacuum polarization

Evidently, the Coulomb potential is associated with the transfer of a single virtual photon. The very simplest approximation to QED, which considers only single-photon exchange between real particles is, roughly speaking, a classical approximation. If, for example, we calculate the scattering cross-section for two electrons using only the single-photon diagram of figure 9.7(*a*), the result obtained in the non-relativistic limit $|\boldsymbol{p}| \ll m$ (and after a substantial amount of algebra) is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{\alpha^2 m^2}{16p^4}\left[\frac{1}{\sin^4(\theta/2)} + \frac{1}{\cos^4(\theta/2)} - \frac{1}{\sin^2(\theta/2)\cos^2(\theta/2)}\right] \qquad (9.86)$$

where $\alpha = e^2/4\pi \approx 1/137$ is the fine structure constant (in SI units, $\alpha = e^2/4\pi\epsilon_0\hbar c$) and, in the centre of mass frame, θ is the scattering angle and p the magnitude of the 3-momentum of each particle. This is a modified version of the classical Rutherford formula, corrections arising from the electrons' being identical spin-$\frac{1}{2}$ particles. Quantum-mechanical corrections, which are all the diagrams containing closed loops, are small in QED, because each photon added to a diagram is attached to a pair of vertices, giving rise to a factor of α. Under some circumstances, however, they can be measured by accurate experiments.

Some, though not all, of these corrections can be regarded as modifications of the photon propagator. For example, figure 9.7(*e*) is obtained from 9.7(*a*) by inserting a closed loop of virtual charged particles, and the same modification can be made to any photon appearing in any diagram. The total effect of such modifications can be represented by replacing each unperturbed photon propagator with the complete propagator, whose first few terms are shown in figure 9.8. After making this replacement, of course, individual diagrams like figure 9.7(*e*) do not appear. By using the complete photon propagator in (9.84), we should obtain a modified Coulomb potential, which describes some of the quantum corrections to classical electrodynamics. This modified potential is said to result from *vacuum polarization*. Picturesquely, the idea is that the electric field of a charged particle polarizes the vacuum, in the sense that the original particle becomes surrounded by a distribution of virtual charged particle-antiparticle pairs, and the net potential is that due to this modified charge distribution.

In momentum space, the contribution to the complete propagator $D_{\mu\nu}(p)$ of the second diagram of figure 9.8 is

$$\mathrm{i}e^2 D_{\mathrm{F}\mu\sigma}(p)\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{\mathrm{Tr}\left[\gamma^\sigma(\not{k}+m)\gamma^\tau(\not{k}+\not{p}+m\right]}{(k^2-m^2+\mathrm{i}\epsilon)[(k+p)^2-m^2+\mathrm{i}\epsilon]}D_{\mathrm{F}\tau\nu}(p) \qquad (9.87)$$

and the set of all diagrams consisting of strings of these loops is, like (9.67) a geometric series. Because the photon propagator always appears inside Feynman diagrams multiplied by e^2, it is useful to consider the quantity $\alpha D_{\mu\nu}(p)$, which must be the sum of a part proportional to $\eta_{\mu\nu}$ and one proportional to $p_\mu p_\nu$. The contribution to the $\eta_{\mu\nu}$ part from the above set of diagrams is

$$\frac{\alpha_0}{[1+\alpha_0 I(p^2)]}\frac{\eta_{\mu\nu}}{(p^2+\mathrm{i}\epsilon)} \tag{9.88}$$

where $I(p^2)$ is an infinite quantity, proportional to the integral in (9.87), and I have added the subscript to α_0 to indicate the need for renormalization.

Our hope is that (9.88) will turn into a finite expression when we rewrite it in terms of the true fine structure constant α, but this raises the question of how α is to be defined. We would expect (and this can be verified *a posteriori*) that modifications of the usual Coulomb potential due to quantum effects should be appreciable only for charged particles separated by a very short distance. The true fine structure constant ought to involve the electronic charge e as measured by macroscopic experimental apparatus, so it can be identified as the coefficient of $1/|\boldsymbol{r}|$ in the large-distance limit of the static potential. As in (9.84), the static potential corresponds to $p^0 = 0$ or $p^2 = -\boldsymbol{p}^2$, and the large-distance limit corresponds to a virtual photon of very large wavelength, which means $\boldsymbol{p} \to 0$. Thus, in the approximation where we use only the diagrams that led to (9.88), we have

$$\alpha = \frac{\alpha_0}{1+\alpha_0 I(0)} \tag{9.89}$$

and (9.88) becomes

$$\frac{\alpha}{1+\alpha[I(p^2)-I(0)]}\frac{\eta_{\mu\nu}}{p^2+\mathrm{i}\epsilon}. \tag{9.90}$$

The difference $I(p^2)-I(0)$ is finite and (again after some lengthy algebra) can be expressed as

$$\begin{aligned} I(p^2) - I(0)& \\ &= -\frac{1}{3\pi}\int_0^1 \mathrm{d}x\left\{\left[1-2Q^{-1}\right]\ln\left[1+x(1-x)Q\right]+2x(1-x)\right\} \\ &\approx -\frac{1}{15\pi}\left(-\frac{p^2}{m^2}\right) \qquad \text{for } |p^2| \ll m^2 \\ &\approx -\frac{1}{3\pi}\ln\left(-\frac{p^2}{m^2}\right) \qquad \text{for } |p^2| \gg m^2 \end{aligned} \tag{9.91}$$

where $Q = -p^2/m^2$. To calculate the static potential, and for some other purposes too, we are interested in negative values of $p^2 = p_0^2 - \boldsymbol{p}^2$, for which Q is positive. By substituting this result into (9.90), we obtain the Fourier transform of the modified Coulomb potential. On carrying out the Fourier transform, we would obtain the modified $V(r)$ itself. The detailed result is a little complicated and not particularly enlightening as it differs significantly from $e^2/4\pi r$ only at extremely short distances. Although we have considered only a single species of charged particle, there will in reality be contributions of the same kind from every species that exists in nature. Clearly, however, the major contribution will be that of the lightest species, namely the electron. The next lightest particle, the muon, is about 200 times heavier and its contribution to the large-distance or low-energy vacuum polarization is much smaller.

9.7.3 The Lamb shift

The modified Coulomb potential which is the spatial Fourier transform of (9.90) will not be exactly proportional to $1/r$. This has a measurable effect upon the atomic spectrum of hydrogen. Readers will recall that in the elementary non-relativistic theory of the hydrogen atom the energy levels are independent of angular momentum and that this fact depends crucially on the form of the Coulomb potential. In a relativistic treatment based on the Dirac equation, the degeneracy is partly lifted by spin-orbit coupling, which leads to the fine-structure splitting, but, for example, the $2S_{1/2}$ and $2P_{1/2}$ levels are still degenerate. If the Coulomb potential is not exactly proportional to $1/r$, then this degeneracy too is lifted. Actually, there are other effects of the loop diagrams of QED which cause a more pronounced 2S-2P splitting than does the vacuum polarization. The measurements of W E Lamb and R C Retherford in 1947 showed the $2P_{1/2}$ level to lie below the $2S_{1/2}$ by an amount corresponding to a frequency $\Delta E/\hbar$ of some 1000 MHz, while a calculation of the vacuum polarization effect alone suggests a shift of about 27 MHz in the opposite direction. However, detailed calculations, including all QED effects and also some nuclear effects, agree with more recent measurements, which give a shift of about 1057.9 MHz, within the experimental accuracy of 0.02 MHz. Since this uncertainty is about a thousand times less than the contribution of the vacuum polarization, the agreement can be taken as confirming the modification of the Coulomb law.

9.7.4 The running coupling constant

The modified Coulomb potential can be interpreted as $V(r) = \alpha(r)/r$, where $\alpha(r)$ is an effective distance-dependent coupling constant. Pictorially, if vacuum polarization is interpreted as a screening of the bare charge of a particle by a cloud of virtual electron-positron pairs, then the apparent charge of the particle depends upon how far into this cloud we have penetrated before measuring it. In Fourier-transformed language, the apparent charge depends upon the wavelength, and thus upon the energy and momentum, of a real or virtual photon that interacts with the charged particle. Using (9.90), we define a *running coupling constant* $\alpha(-p^2)$ by

$$\alpha(-p^2) = \frac{\alpha}{1+\alpha[I(p^2) - I(0)]}. \tag{9.92}$$

There are several important theoretical issues associated with this running coupling constant. In the first place, there is a close link with the process of renormalization. Instead of using the true fine-structure constant, we could in principle define a renormalized coupling constant in terms of the value of (9.88) at $p^2 = -\mu^2$, μ being an arbitrary parameter as in (9.73). Then, in (9.89) and (9.90), $I(0)$ would be replaced by $I(-\mu^2)$. We easily find that

$$\alpha(\mu^2) = \frac{\alpha}{1 + \alpha[I(-\mu^2) - I(0)]} \tag{9.93}$$

which is the same equation as (9.92). This facet of renormalization can be developed more thoroughly. The resulting machinery has come to be known as the *renormalization group*, and I shall explore one of its uses in chapter 11.

The existence of the running coupling constant can be taken to mean that the effective strength of electromagnetic interactions varies with energy. The variation is appreciable only when $(-p^2) \gg m^2$, and in that limit we have

$$\alpha(-p^2) \approx \alpha \left[1 - \frac{\alpha}{3\pi} \ln\left(-\frac{p^2}{m^2} \right) \right]^{-1} . \tag{9.94}$$

At the energies of a TeV or so ($1\,\text{TeV} = 10^{12}\,\text{eV}$) that are accessible in modern particle accelerators, $\alpha(-p^2)$ has increased by only about 2% from its zero-energy value α. On the other hand, we see that $\alpha(-p^2)$ becomes infinite when $(-p^2) = m^2 \exp(411\pi)$. This energy is so vast as to be irrelevant to any conceivable experiment, but there is cause for concern on theoretical grounds. The pole in (9.90) at $p^2 = 0$ is, as we know, associated with the existence of real photons of zero mass. An infinite value of the running coupling constant would seem to imply the existence of a particle with imaginary mass M given by $M^2 = -m^2 \exp(411\pi)$, sometimes referred to as the *Landau ghost*. This would be a *tachyonic*, or faster-than-light particle, since $v^2/c^2 = 1 - M^2/E^2$ at energy E. Such particles are generally believed to be impossible, so the Landau ghost seems to indicate some fundamental flaw in QED. A related problem is that the infinite constant $I(0)$ in (9.89) is positive. This appears to mean that there is no positive value of α_0 (and therefore no real value of e_0, even zero or infinity, for which the renormalized α is non-zero. If every permissible value of α_0 leads to α being zero, then the theory is in fact non-interacting (we are not allowed to set $\alpha = 1/137$) and is said to be *trivial*. This question is somewhat confused, because the arguments are based on approximations of one kind or another and the bare coupling α_0 has no direct physical meaning. There is no doubt that perturbative QED is an excellent theory of electromagnetism at experimentally accessible energies, but many believe that it would break down at sufficiently high energies and, indeed, that it ultimately makes sense only when embedded in a more complete theory.

9.7.5 Anomalous magnetic moments

A charged, spinning particle might be expected to possess a magnetic dipole moment, and so it does. An extremely accurate test of QED is provided by measurements of the magnetic moments of the electron and muon. To see how these are calculated, it is helpful first to study the non-relativistic limit of the Dirac equation (8.13) which, for an electron of charge $-e$, reads

$$(\mathrm{i}\not{\partial} + e\not{A} - m)\psi = 0. \tag{9.95}$$

When the kinetic energy is much smaller than the rest energy m, we can, approximately, derive a Schrödinger equation from this. We first multiply on the left by γ^0 to give

$$\mathrm{i}\frac{\partial\psi}{\partial t} = \left(-\mathrm{i}\gamma^0\gamma^i\partial_i - eA_0 - e\gamma^0\gamma^i A_i + m\gamma^0\right)\psi \tag{9.96}$$

and note that, in the standard representation of the γ matrices, we have

$$\gamma^0\gamma^i = \begin{pmatrix} 0 & \sigma^i \\ \sigma^i & 0 \end{pmatrix} \qquad \text{and} \qquad \gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}. \tag{9.97}$$

When m is large compared with the kinetic energy, the most rapid time dependence of ψ is in a factor $\exp(-\mathrm{i}mt)$. For a free, positive-energy particle in its rest frame, the solution is $\exp(-\mathrm{i}mt)$ multiplied by one of the spinors (7.68). For small kinetic and electromagnetic energies, therefore, we anticipate a solution of the form

$$\psi = \mathrm{e}^{-\mathrm{i}mt}\begin{pmatrix} \chi \\ \theta \end{pmatrix} \tag{9.98}$$

where χ and θ are two-component spinors and θ is small. On substituting this into (9.96), we obtain two coupled equations for χ and θ:

$$\mathrm{i}\frac{\partial\chi}{\partial t} = -\sigma^i(\mathrm{i}\partial_i + eA_i)\theta - eA^0\chi \tag{9.99}$$

$$\mathrm{i}\frac{\partial\theta}{\partial t} = -\sigma^i(\mathrm{i}\partial_i + eA_i)\chi - eA^0\theta - 2m\theta. \tag{9.100}$$

When m is large and θ is small, the solution to (9.100) is approximately

$$\theta \approx -\frac{1}{2m}\sigma^i(\mathrm{i}\partial_i + eA_i)\chi \tag{9.101}$$

and by substituting this into (9.99) we find

$$\mathrm{i}\frac{\partial\chi}{\partial t} = -\frac{1}{2m}\sigma^i\sigma^j(\boldsymbol{\nabla} + \mathrm{i}e\boldsymbol{A})^i(\boldsymbol{\nabla} + \mathrm{i}e\boldsymbol{A})^j\chi - e\Phi\chi \tag{9.102}$$

where $\Phi = A^0$ is the electric potential.

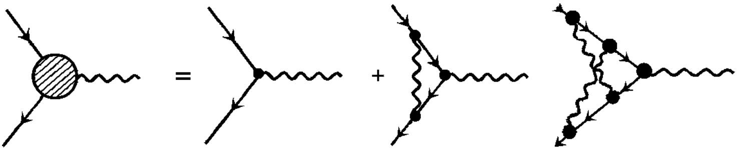


Figure 9.9. The effective electron–photon vertex which gives rise to an anomalous magnetic moment.

Now, the Pauli matrices satisfy the identity

$$\sigma^i\sigma^j = \delta^{ij} + \mathrm{i}\epsilon^{ijk}\sigma^k \tag{9.103}$$

which leads to the final result

$$\mathrm{i}\frac{\partial\chi}{\partial t} = \left[-\frac{1}{2m}(\nabla + \mathrm{i}e\boldsymbol{A})^2 - e\Phi + \frac{e}{m}\frac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{B}\right]\chi. \tag{9.104}$$

The first two terms on the right-hand side give the usual Schrödinger equation for a particle of charge $-e$ in an electric potential Φ and magnetic vector potential $\boldsymbol{A}$. The last term represents the interaction of a magnetic moment $\boldsymbol{\mu} = (-e/m)(\frac{1}{2}\boldsymbol{\sigma})$ with the magnetic field $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. Since the spin angular momentum operator is $\boldsymbol{s} = \frac{1}{2}\boldsymbol{\sigma}$, we have

$$\boldsymbol{\mu} = -g_{\mathrm{s}}\mu_{\mathrm{B}}\boldsymbol{s} \tag{9.105}$$

where $\mu_{\mathrm{B}} = e/2m$ is the *Bohr magneton* and $g_{\mathrm{s}} = 2$. This is a somewhat surprising prediction of the Dirac equation, because the corresponding g factor for orbital angular momentum is 1.

Experimentally, this prediction is approximately verified for electrons and muons, but there is a correction of about 1% arising from higher-order quantum effects in QED. The way this comes about is quite similar to the modification of the Coulomb potential by vacuum polarization. In (9.95), the middle term is $e\gamma^\mu A_\mu\psi$, and the γ^μ is the same as the one that appears in the QED vertex (9.53). Now consider again the 4-point Green function which is the sum of the diagrams of figure 9.7, along with many others. The total effect of diagrams (*a*), (*b*) and an infinite set of similar ones can be obtained by keeping just (*a*), but replacing the γ^μ in its upper vertex with an effective vertex Γ^μ, which is the sum of a series of diagrams whose first few terms are shown in figure 9.9. In the same way, any Green function can be expressed as a sum of 'skeleton' diagrams, in which each vertex is Γ^μ and diagrams such as figure 9.7(*b*) do not appear. Thus, the vertex Γ^μ represents the net effect of higher-order corrections on the interaction between an electron and a photon, as suggested by figure 9.9. Essentially, the *anomalous magnetic moment* is calculated by replacing γ^μ with Γ^μ in the previous calculation, but the technical details are a little complicated. The anomaly is defined by $a = (g_{\mathrm{s}} - 2)/2$ and its lowest-order contribution is

$\alpha/2\pi$. The best theoretical and experimental values for the electron anomaly are

$$a_{\mathrm{th}} = (1\ 159\ 652.2 \pm 0.2) \times 10^{-9}$$
$$a_{\mathrm{exp}} = (1\ 159\ 652.188 \pm 0.004) \times 10^{-9}.$$

As a matter of fact, theoretical calculations have been carried out which are rather more accurate than the quoted uncertainty suggests. Most of this uncertainty is the experimental uncertainty in the value of α which has to be substituted into the calculated formula. For muons, there is similar agreement between theory and experiment, although the accuracy of each is somewhat less. Moreover, there are strong- and weak-interaction corrections to the magnetic moment as calculated using QED alone. For the electron, these corrections are no bigger than the uncertainty; for the muon they turn out to be more important, and must be included to obtain agreement with the measured value.

For the proton and neutron, the g factors found from the Dirac equation are 2 and 0 respectively, but they are found experimentally to be approximately 5.58 and -3.82. The reason for these large discrepancies is that the Dirac equation applies to point particles. The experimental values for the various magnetic moments may be taken as evidence that, whereas the electron and muon are truly fundamental particles, the proton and neutron have an internal structure, being composed of more elementary constituents, the quarks. Although theoretical models of the quark structure of nucleons are by no means as accurate as QED, the observed magnetic moments can be reasonably well accounted for on this basis.

Exercises

9.1. In many contexts, Green functions of various kinds are encountered as a means of solving differential equations. If $\phi_0(x)$ is a solution of the Klein–Gordon equation $(\Box + m^2)\phi_0 = 0$, show that a solution of the equation $(\Box + m^2)\phi(x) = j(x)$ is given by

$$\phi(x) = \phi_0(x) - \int \mathrm{d}^4y\, G_{\mathrm{F}}(x-y)j(y).$$

9.2. In equation (7.11), denote the positive-energy part of $\phi(x)$ by $\phi_a(x)$ and the negative-energy part by $\phi_c^*(x)$. Show that

$$\int \mathrm{d}^3x\, G_{\mathrm{F}}(\boldsymbol{x}' - \boldsymbol{x}, t' - t)\frac{\overleftrightarrow{\partial}}{\partial t}\phi(\boldsymbol{x}, t)$$
$$= -\theta(t' - t)\phi_a(\boldsymbol{x}', t') + \theta(t - t')\phi_c^*(\boldsymbol{x}', t')$$

where θ is the step function (see appendix A). Can you justify Feynman's description of an antiparticle as 'a particle travelling backwards in time'?

9.3. Write down an expression for the time-ordered product of two bosonic or fermionic field operators, using the step functions $\theta(x^0 - y^0)$ and $\theta(y^0 - x^0)$ to distinguish the two time orderings. Use Cauchy's theorem to show that the step function can be represented as

$$\theta(t - t') = \lim_{\epsilon \to 0} \frac{1}{2\pi \mathrm{i}} \int_{-\infty}^{\infty} \mathrm{d}\omega \frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{\omega - \mathrm{i}\epsilon}.$$

By expressing the free field operators in terms of creation and annihilation operators, verify the expressions (9.40) and (9.44) for the scalar and spinor propagators.

9.4. The symbol $\Box^{-1}$ means that if $\Box A = B$ then $A = \Box^{-1} B$. For example, $\Box^{-1} \exp(\mathrm{i}k \cdot x) = -\exp(\mathrm{i}k \cdot x)/k^2$. The transverse and longitudinal projection operators $T_{\mu\nu}$ and $L_{\mu\nu}$ are defined by $T_{\mu\nu} = \eta_{\mu\nu} - \partial_\mu \partial_\nu \Box^{-1}$ and $L_{\mu\nu} = \partial_\mu \partial_\nu \Box^{-1}$.

Show that (a) $L_{\mu\nu} + T_{\mu\nu} = \eta_{\mu\nu}$; (b) $L_{\mu\sigma} L^{\sigma}_{\ \nu} = L_{\mu\nu}$; (c) $T_{\mu\sigma} T^{\sigma}_{\ \nu} = T_{\mu\nu}$; (d) $L_{\mu\sigma} T^{\sigma}_{\ \nu} = T_{\mu\sigma} L^{\sigma}_{\ \nu} = 0$. Solve (9.63) by expressing the differential operator in terms of these projection operators and by expressing $D_{\mathrm{F}\mu\nu}(x - y)$ in terms of projection operators acting on $\delta(x - y)$. (For this purpose, set $\epsilon = 0$.)

9.5. A charged particle of mass m undergoes an electromagnetic scattering process, emitting a virtual photon that subsequently interacts with another particle. If p^μ and p'^μ are the initial and final 4-momenta of the particle ($p^2 = p'^2 = m^2$), then the 4-momentum of the virtual photon is $q^\mu = p^\mu - p'^\mu$. Show that $q^2 \le 0$. [Hint: consider the frame of reference in which $\boldsymbol{p}' = -\boldsymbol{p}$.]

9.6. (a) In equation (9.95), take $A_\mu(x)$ to be a real function, representing an externally applied electromagnetic field. By considering the charge conjugate of this equation, show that particles and antiparticles have opposite electric charges.

(b) Now consider the proposition that charge conjugation is a symmetry of nature, in the sense that a state in which all particles are replaced with their antiparticles is indistinguishable from the original state. (This is true of a universe with only electromagnetic forces, but not of a universe in which there are weak interactions as well.) Consider $A_\mu(x)$ to be a field operator. Then the charge conjugate of (9.95) should be equivalent to *exactly the same equation*, but with both ψ and A_μ replaced by their charge conjugates. Show that $A^{\mathrm{c}}_\mu = -A_\mu$.

(c) To get the correct answer for (a), you should *not* have replaced A_μ with $-A_\mu$. Convince yourself that (a) and (b) are consistent by considering how the electromagnetic fields produced by a given distribution of charged particles are affected by reversing the charges of these particles without changing their state of motion, and whether, in (a), the charges of *all* relevant particles were reversed.

Chapter 10

Equilibrium Statistical Mechanics

When we deal with systems containing many particles, it soon becomes essential to adopt statistical methods of analysis. To a large extent, statistical mechanics has been developed with a view to studying condensed matter systems, such as solids and fluids, upon which controlled laboratory experiments can be performed. In some cases, the quantum-mechanical properties of the constituent particles are crucial. This is true, for example, when we study the properties of electrons in metals or semiconductors, or of superfluid helium. In other cases, it is sufficient to treat the constituent particles according to classical mechanics, although it may still be necessary to determine their properties, such as the forces which act between them, from the underlying quantum theory. The properties of most normal fluids and many magnetic properties of solid materials, for example, can be adequately and conveniently treated by classical methods.

There are, moreover, important connections between statistical mechanics and the relativistic field theories that have been our concern in previous chapters. Indeed, the entire history of quantum mechanics and quantum field theory might be said to have started with Planck's attempts to understand black-body radiation in terms of statistical mechanics. The most obvious connection is that it may be necessary to consider the behaviour of large assemblages of high-energy particles, whose proper description is in terms of quantum field theory. Black-body radiation is a case in point, although it can be understood without the full machinery of field theory. Other examples are the hot, dense gases found, it is thought, in the cores of some stars or in the early universe and, perhaps, small amounts of hot matter formed in high-energy collisions of heavy ions. At the mathematical level, there are close formal similarities between the thermal averages of statistical mechanics and the functional integral methods of quantum field theory, which I shall discuss towards the end of this chapter. The recognition of these similarities has proved enormously fruitful. For example, the methods of quantum field theory have shed considerable light on certain problems in condensed matter physics, especially those involving phase transitions, as we shall see in the next chapter, while techniques developed originally for statistical mechanics provide alternative methods of approximation in relativistic field theories, when perturbation theory is not applicable.

In this book, I shall consider, for the most part, only *equilibrium statistical mechanics*. The assumption of thermal equilibrium, that is, of a state in which all macroscopic properties of the system have settled down to constant values, leads to great simplifications, provided we accept that the measured values of these quantities are to be compared with suitably weighted averages over the microscopic states of our theoretical model system. For we then have only to establish what weight should be attached to a given state and are absolved from considering how the system passes from one state to another. The mathematical foundations of statistical mechanics have been developed rather more fully for classical systems than for quantum-mechanical ones. I shall begin by considering the kinds of justification that have been suggested for the use of particular statistical weight functions for classical systems and then examine the relationship between statistical mechanics and thermodynamics. Finally, I shall describe the adaptation of these ideas to quantum mechanics and quantum field theory.

10.1 Ergodic Theory and the Microcanonical Ensemble

It will probably strike readers as intuitively obvious that macroscopic measurements generally yield some kind of average value of the measured quantity. This is because of the limited resolution of our measuring apparatus, but there are at least two different aspects to this, both of which are called upon to justify different theoretical steps. Consider, for example, a largeish amount of a gas in a transparent container. Suppose, for the sake of argument, that we know, with negligible error, the total mass of gas and the volume of the container. Then the ratio of the two gives us a value for the overall density. By passing a beam of light through the container, we can measure the refractive index, and hence the density, of that region of the gas that the beam intersects. There are two reasons for expecting the density measured in this way to coincide with the overall density. One is that the measurement process takes much longer than the timescales which characterize the microscopic motions (for example, the mean time between two collisions of a single particle or the time taken for a particle to cross the beam). Therefore, although the number of particles in the volume defined by the light beam fluctuates with time, we would expect the measured density to be a *long time average* of instantaneous densities and, further, that this average should coincide with the overall density. The second reason is that, even though the volume defined by the beam may be only a small part of the total volume, it will normally contain a large number of particles. Averaged over all possible configurations of the particles, the density should certainly be equal to the overall density, and probability theory tells us to expect relative fluctuations about this average that depend inversely on the square root of the mean number of particles. Because our measurement is *coarse grained*, in the sense that it probes distances much greater than the average separation of two particles, we would expect even an instantaneous measurement to give a value very close to the average.

The statistical description of systems in thermal equilibrium is based on the idea that the measured value of a quantity is a long-time average. We further assume that, during the time taken to perform the measurement, the system passes through a sequence of instantaneous states that is representative of the whole set of states available to it. In classical mechanics, the instantaneous state of a system can be represented as a point in *phase space*. For a system of N particles, phase space Γ is the $6N$-dimensional manifold (discussed from a geometrical point of view in §3.7) whose points correspond to the values of the $3N$ coordinates and $3N$ momenta. For the moment, it will be convenient to lump the coordinates and momenta together into a $6N$-dimensional coordinate X. A weighted average of a quantity $f(X)$ is of the form

$$\langle f\rangle_t = \int_\Gamma \mathrm{d}^{6N}X\,\rho(X,t)f(X) \tag{10.1}$$

where $\rho(X,t)$ is a probability density for finding the system in a state close to X at time t. The probability density can be visualized in terms of a *Gibbs ensemble* of very many identical systems, $\rho(X,t)\mathrm{d}^{6N}X$ being the fraction of these whose state at time t is in the phase-space volume element $\mathrm{d}^{6N}X$ containing X.

An equation governing the rate of change of the probability distribution with time can be deduced from Hamilton's equations (3.16). In fact, we have already derived this equation, namely the Liouville equation (3.22), for the particular distribution (3.20). To show that the same equation is valid for any other distribution, we consider the points representing members of the ensemble as a 'probability fluid' in phase space. The current density of this fluid has components $j_i = \dot{X}_i\rho(X,t)$ and, since we are not going to change the probability artificially by adding or removing systems from the ensemble, the equation of continuity must hold:

$$\frac{\partial}{\partial t}\rho(X,t) = -\sum_{i=1}^{6N}\frac{\partial}{\partial X_i}[\dot{X}_i\rho(X,t)]. \tag{10.2}$$

From Hamilton's equations, we find

$$\sum_{i=1}^{6N}\frac{\partial\dot{X}_i}{\partial X_i} = \sum_{i=1}^{3N}\left(\frac{\partial\dot{q}_i}{\partial q_i}+\frac{\partial\dot{p}_i}{\partial p_i}\right) = \sum_{i=1}^{3N}\left(\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i}-\frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i}\right) = 0 \tag{10.3}$$

and therefore

$$\frac{\partial}{\partial t}\rho(X,t) = -\sum_{i=1}^{6N}\dot{X}_i\frac{\partial}{\partial X_i}\rho(X,t) = -\mathrm{i}\mathcal{H}\rho(X,t) \tag{10.4}$$

where $\mathcal{H}$ is the Liouville operator defined in (3.19). This is the *Liouville equation*. It gives the rate of change of the probability density at a fixed point in phase space.

We could also fix our attention on a particular member of the ensemble, whose state is $X(t)$, and ask how the probability density in its neighbourhood, $\rho(X(t), t)$ changes with time. The answer is

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(X(t), t) = \frac{\partial}{\partial t}\rho(X(t), t) + \sum_{i=1}^{6N} \dot{X}_i \frac{\partial}{\partial X_i}\rho(X(t), t) = 0. \tag{10.5}$$

This result, known as *Liouville's theorem*, is usually described by saying that the probability density behaves as an incompressible fluid. It does not, however, imply that ρ has a uniform value over that part of phase space where it is non-zero, as would be true for an ordinary incompressible fluid.

For a system in equilibrium, all averages of the form (10.1) should be constant in time, which means that $\partial\rho/\partial t = 0$. According to (10.4), this will be true if ρ depends on X only through quantities whose Poisson brackets with the Hamiltonian H are zero, which are conserved quantities. For simplicity, I shall assume that the only relevant conserved quantity is the energy. The probability density that describes a system in equilibrium depends, as we shall see, on how the system is allowed to interact with its environment. Once this interaction is specified, it is quite straightforward to construct the appropriate probability density. Ideally, however, we would like to have some reassurance on several points. First, we would like to know whether the ensemble average (10.1) is indeed equal to the long-time average which, by hypothesis, corresponds to an experimental measurement. If so, we would like to be sure that the time-independent probability density we have constructed is unique, for if more than one could be found we would have no good reason for preferring any particular one. Finally, we would like to understand theoretically why a system that starts in a non-equilibrium state usually does settle down into a state of thermal equilibrium. The theory that tries to answer these questions in a mathematically rigorous manner is called *ergodic theory*. It is unfortunately true that, while many elegant mathematical results have been obtained, the effort required to derive them is out of all proportion to their practical utility in applications to actual physical systems. I shall therefore not attempt to do more than convey the flavour of what is involved.

We consider a system that is completely isolated from its environment. It is therefore *closed*, which means that no particles enter or leave it, and *isoenergetic*, which means that its energy is fixed at a definite value E. The probability density must be zero except on the $(6N-1)$-dimensional surface where $H(X) = E$. A candidate for the equilibrium probability density, which depends on the phase-space point X only through $H(X)$, is

$$\rho_{\mathrm{micro}}(X, E) = \frac{\delta[H(X) - E]}{\Sigma(E)} \tag{10.6}$$

where, to ensure the correct normalization,

$$\Sigma(E) = \int \mathrm{d}^{6N}X\, \delta[H(X) - E]. \tag{10.7}$$

The Gibbs ensemble corresponding to this probability density is called the *microcanonical ensemble*. It is uniformly distributed over the constant energy surface.

The microcanonical ensemble is likely to be relevant to experimental observations if the averages we calculate with it are equal to the corresponding long-time averages. A system is said to be *ergodic* if, for any smooth function $f(X)$,

$$\int_{\Gamma} \mathrm{d}^{6N}X \, \rho_{\text{micro}}(X) f(X) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}t \, f(X(t)) \tag{10.8}$$

and if this is true for almost all starting points $X(t = 0)$ for the trajectory on the right-hand side. The phrase 'almost all' has the mathematical sense of 'except on a set of zero measure', which means that the set of exceptional starting points makes no contribution to the ensemble average on the left. The way this might come about is as follows. Imagine the constant-energy surface to be divided into small cells. In the course of its motion over a very long time, the point $X(t)$ representing an ergodic system will pass through every cell, provided that we wait long enough, and the fraction of time that it spends in each cell is equal to the weight of that cell in the ensemble average. This is true for any cells of finite size, however small, so the trajectory will eventually pass arbitrarily close to any point of the energy surface. The stronger statement that it will eventually pass *through* every point is actually not true. The application of the microcanonical ensemble to averages in thermal equilibrium is justified by the *ergodic theorem* due to G D Birkhoff and A I Khinchin, which states that, for an ergodic system, the microcanonical ensemble is the only time-independent probability density on the energy surface. The converse, that a system for which the only time-independent distribution is the microcanonical one is ergodic, is also true. The drawback of this approach lies in the extreme difficulty of proving that any system of real physical interest actually *is* ergodic. One such proof, given by Y Sinai, applies to a gas of hard spheres; that is, to a gas of spherical molecules which do not deform or penetrate each other, but exert no other forces. Given that this admittedly idealized model system is ergodic, we might expect that other, more realistic models would also have this property.

Although ergodicity ensures that the microcanonical ensemble correctly describes thermal equilibrium, it does not ensure that an isolated system will eventually settle into equilibrium if it starts in some other state. In other words, a Gibbs ensemble which initially does not have the uniform microcanonical distribution over the energy surface will not necessarily approach such a distribution with the passage of time. On the face of it, indeed, it seems unlikely that this could ever happen. From (10.5), we know that the density in the neighbourhood of any particular member of the ensemble is constant, and therefore any initial inhomogeneities in $\rho(X)$ cannot be smoothed out with time, although they will move around the energy surface. The kind of thing that might happen is illustrated schematically in figure 10.1, where ρ is zero, except in the

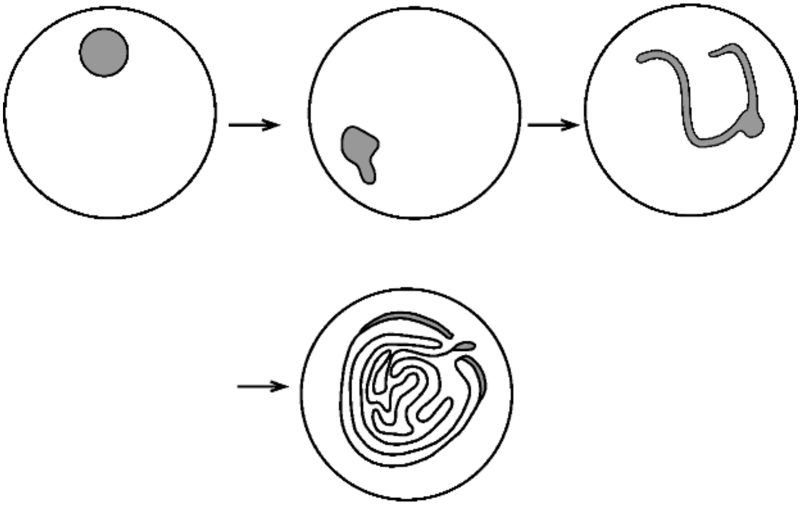


Figure 10.1. Schematic illustration of the evolution in time of the phase-space probability density of a mixing system. The probability density is non-zero only in the shaded region, whose area is constant.

shaded region. The fraction of the energy surface where ρ is non-zero is constant in time, but the shape of this region may evolve in a complicated way, developing strands which spread out over the entire energy surface. If the surface is divided into small cells, and we define a *coarse-grained* probability density by averaging over each cell, then this coarse-grained probability density may well become uniform. Since our experimental measurements are in any case coarse grained, the actual probability density would, for practical purposes, become indistinguishable from the microcanonical one, because we would only want to average functions $f(X)$ that vary very little within a coarse-graining cell.

This kind of behaviour is somewhat analogous to the mixing of two immiscible liquids, such as oil and water, stirred together in a container to produce a mixture that is homogeneous in the coarse-grained sense. Systems whose trajectories in phase space lead to this kind of development of a probability density are called *mixing*. There is, of course, a precise mathematical definition, but we shall not be making any use of it. It can be shown that all mixing systems are also ergodic, but the converse is not true. The hard-sphere gas was in fact shown by Sinai to be mixing.

A simple example of the use of the microcanonical ensemble is provided by an ideal gas with Hamiltonian

$$H = \sum_{i=1}^{N} \frac{1}{2m} \boldsymbol{p}_i^2 \tag{10.9}$$

confined to a volume V. The area of the energy surface $\Sigma(E)$ can be expressed as

$$\begin{aligned}\Sigma(E) &= \int \mathrm{d}^{3N}p\,\mathrm{d}^{3N}x\,\delta\left(E - \frac{1}{2m}\sum_{i=1}^{3N} p_i^2\right) \\ &= \frac{\partial}{\partial E}\int \mathrm{d}^{3N}p\,\mathrm{d}^{3N}x\,\theta\left(E - \frac{1}{2m}\sum_{i=1}^{3N} p_i^2\right) \qquad (10.10)\end{aligned}$$

where $\theta(E - H)$ is the step function. The integral over coordinates gives V^N and the momentum integral is the volume of a $3N$-dimensional sphere of radius $(2mE)^{1/2}$, which can be evaluated as in appendix A. The final answer is

$$\Sigma(E) = \frac{V^N(2\pi m)^{3N/2}E^{(3N/2)-1}}{(\frac{3}{2}N - 1)!} \qquad (10.11)$$

and we shall see shortly that it is related to the entropy of the gas.

10.2 The Canonical Ensemble

If our system is allowed to exchange heat energy with its surroundings, we need a somewhat different statistical description. So long as we restrict ourselves to equilibrium conditions, we need not be very precise about the mechanism that allows this exchange to take place. The simplest course is to suppose that the surroundings constitute a *heat bath*. Ideally, the heat bath is an infinite system, which can exchange finite amounts of energy with the system of interest without any change in its own properties. Experimentally, this situation can be accurately simulated by using thermostatic feedback techniques. Normally, we describe these as techniques for maintaining a constant temperature, but we have yet to establish a precise notion of temperature within statistical mechanics. We shall still take the total number of particles in the system to be fixed, in which case we are dealing with a *closed isothermal system*. The Gibbs ensemble for such a system is called the *canonical ensemble* and our first objective is to find the appropriate probability density $\rho_{\mathrm{can}}(X)$. The question of what this probability density should be has not been investigated with the same degree of mathematical rigour as for the microcanonical ensemble, but the following simple argument produces what is universally accepted as the correct answer.

Consider two systems, A and B, which are both in equilibrium with the same heat bath but do not interact directly with each other. Individually, they have probability densities $\rho_{\mathrm{can}}(X_{\mathrm{A}})$ and $\rho_{\mathrm{can}}(X_{\mathrm{B}})$, which depend on the coordinates and momenta only through $H_{\mathrm{A}}(X_{\mathrm{A}})$ and $H_{\mathrm{B}}(X_{\mathrm{B}})$ respectively. Equally, we can regard A and B as a single system AB, with Hamiltonian $H_{\mathrm{AB}}(X_{\mathrm{AB}}) = H_{\mathrm{A}}(X_{\mathrm{A}}) + H_{\mathrm{B}}(X_{\mathrm{B}})$, whose probability density $\rho_{\mathrm{can}}(X_{\mathrm{AB}})$ depends only on H_{AB}.

Since A and B do not interact, their probability densities should be independent, and the joint probability density is

$$\rho_{\rm can}(H_{\rm AB}) = \rho_{\rm can}(H_{\rm A} + H_{\rm B}) = \rho_{\rm can}(H_{\rm A})\rho_{\rm can}(H_{\rm B}). \tag{10.12}$$

This relation determines the form of $\rho_{\rm can}$ up to a single parameter. For a function of a single variable, $f(x)$, which has the property $f(x+y) = f(x)f(y)$, we can first deduce that $f(0) = 1$ by setting $x = y = 0$. Then, by choosing y to be a small increment δx and defining $\beta = -f'(0)$, we obtain the differential equation $\mathrm{d}f(x)/\mathrm{d}x = -\beta f(x)$. Since $f(0) = 1$, the unique solution is $f(x) = \exp(-\beta x)$. In (10.12), the analogue of x is the function $H(X)$, and this allows some extra freedom in the normalization. It is easy to see that the normalized probability density that satisfies (10.12) is

$$\rho_{\rm can}(X, \beta) = \mathrm{e}^{-\beta H(X)} \left[\int \mathrm{d}^{6N} X \, \mathrm{e}^{-\beta H(X)} \right]^{-1}. \tag{10.13}$$

The undetermined constant β is the same for any system in contact with the same heat bath, so it must be a property of the heat bath itself. Thermodynamically, the only relevant property is its temperature. Thus, β must be a function of temperature, and we can clearly relate it to the ideal gas scale of temperature by taking the system to be an ideal gas.

For a gas or liquid consisting of N identical molecules, we define the *canonical partition function* $Z_{\rm can}(\beta, V, N)$ in terms of the normalizing factor in (10.13) by

$$Z_{\rm can}(\beta, V, N) = \frac{1}{h^{3N} N!} \int \mathrm{d}^{6N} X \, \mathrm{e}^{-\beta H(X)}. \tag{10.14}$$

By including the $1/N!$, we get a sum over all distinct states of the system, counting any two states that differ only by the interchange of a pair of particles as indistinguishable. The factor h^{-3N} has no physical significance and is included as a matter of theoretical convenience to make $Z_{\rm can}$ dimensionless. The constant h is arbitrary, but must have the dimensions of an action. It is convenient to take it to be Planck's constant, because this allows a direct comparison to be made between corresponding classical and quantum-mechanical systems. Many quantities of thermodynamic interest can be expressed as derivatives of the partition function. In particular, the average internal energy U is evidently given by

$$\begin{aligned} U(\beta, V, N) &= \int \mathrm{d}^{6N} X \, H(X) \mathrm{e}^{-\beta H(X)} \left[\int \mathrm{d}^{6N} X \, \mathrm{e}^{-\beta H(X)} \right]^{-1} \\ &= -\frac{\partial}{\partial \beta} \ln Z_{\rm can}(\beta, V, N). \end{aligned} \tag{10.15}$$

For an ideal monatomic gas, we easily obtain

$$Z_{\rm can}(\beta, V, N) = \frac{V^N}{N!} \left(\frac{2\pi m}{\beta h^2} \right)^{3N/2} \tag{10.16}$$

and the internal energy is found to be $U = 3N/2\beta$. For this gas, an elementary kinetic argument (see exercise 10.2) shows that the pressure is related to the internal energy by $pV = \frac{2}{3}U$, so we have $pV = N/\beta$. The ideal-gas scale of temperature is defined by the equation of state $pV = Nk_{\mathrm{B}}T$, where, in SI units, $k_{\mathrm{B}} = 1.380\,54 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is Boltzmann's constant and T is the absolute temperature, so we identify

$$\beta = 1/k_{\mathrm{B}}T. \tag{10.17}$$

10.3 The Grand Canonical Ensemble

A system which can exchange both heat energy and particles with its surroundings is called an *open isothermal system.* Exactly what this means depends to some extent on the particular physical situation we want to investigate. Most straightforwardly, we can think of a very large homogeneous system, within which we draw an imaginary boundary enclosing a small part of the whole, which still contains a very large number of particles. Our earlier example of a light beam intersecting a large container of gas would be a case in point. The small subsystem constitutes 'the system' while the remainder of the original large system acts as an (ideally infinite) heat bath and particle reservoir. The Gibbs ensemble that describes an open isothermal system is the *grand canonical ensemble.*

The grand canonical probability density allows for the possibility of the system's containing any number of particles. It must have the general form

$$\rho(X) = g_N \exp[-\beta H_N(X)] \left[\sum_{N=0}^{\infty} g_N \int \mathrm{d}^{6N}X \, \exp[-\beta H_N(X)]\right]^{-1} \tag{10.18}$$

where H_N is the Hamiltonian of the system when it contains exactly N particles and g_N is related to the probability that it does contain N particles. This probability is obtained by integrating over the coordinates and momenta that the N particles might have:

$$P_N = g_N \int \mathrm{d}^{6N}X \, \exp[-\beta H_N(X)] \left[\sum_{N=0}^{\infty} g_N \int \mathrm{d}^{6N}X \, \exp[-\beta H_N(X)]\right]^{-1}. \tag{10.19}$$

If a particular particle can find itself, with equal probability, anywhere in the system or reservoir, and the reservoir is very much larger than the system, then the probabilities P_N should form a Poisson distribution

$$P_N = \frac{\bar{N}^N \mathrm{e}^{-\bar{N}}}{N!} \tag{10.20}$$

where $\bar{N}$ is the average number of particles in the system.

In the case of non-interacting particles, the N-particle Hamiltonian is just the sum of single-particle Hamiltonians, and

$$\int \mathrm{d}^{6N}X\, \exp[-\beta H_N(X)] = \left[\int \mathrm{d}^3x\mathrm{d}^3p\, \exp[-\beta H_1(x,p)]\right]^N = (h^3 Z_1)^N \tag{10.21}$$

where Z_1 is the canonical partition function for a single particle. The two expressions (10.19) and (10.20) are then consistent if we set

$$g_N = \frac{1}{N!}\left(\frac{\bar{N}}{h^3 Z_1}\right)^N . \tag{10.22}$$

In general, the grand canonical probability density is defined as

$$\begin{aligned}&\rho_{\mathrm{gr}}(N, X, \beta, \mu)\\ &\quad= \frac{z^N}{h^{3N}N!}\exp[-\beta H_N(X)]\left[\sum_{N=0}^{\infty}\frac{z^N}{h^{3N}N!}\int \mathrm{d}^{6N}X\exp[-\beta H_N(X)]\right]^{-1}\end{aligned} \tag{10.23}$$

where the *fugacity* z is

$$z = \mathrm{e}^{\beta\mu} \tag{10.24}$$

and μ is called the *chemical potential.* The chemical potential is taken to be a property of the particle reservoir and so, while it controls the average number $\bar{N}$ of particles in the system, it is independent of the number N that characterizes a particular configuration of the system.

From the derivation of (10.22), it is clear that the general expression (10.23) for the grand canonical probability density is strictly valid only when the integral $Y_N = \int \mathrm{d}^{6N}X\, \exp[-\beta H_N(X)]$ can be written as Y^N, where Y is a quantity independent of N. This is usually not true when particles interact, but it is an excellent approximation when we consider a large system and interactions that are appreciable only over a distance which is small compared with the dimensions of the system. In that case, we can divide the volume of the system into a large number of cells, each of a size greater than the range of interactions, and ignore interactions between particles in different cells. The integral Y then factorizes into a product of single-cell terms, and the number of these terms is proportional to the number of particles in the system. Finally, since the relative fluctuations in the number of particles in a large system are small, only those terms in (10.23) for which N is large will be important.

The *grand canonical partition function* is defined as the normalizing denominator in (10.23):

$$Z_{\mathrm{gr}}(\beta, V, \mu) = \sum_{N=0}^{\infty}\exp(\beta\mu N)Z_{\mathrm{can}}(\beta, V, N)$$

$$= \sum_{N=0}^{\infty} \frac{1}{h^{3N} N!} \exp(\beta\mu N) \int \mathrm{d}^{6N} X \exp[-\beta H_N(X)]. \quad (10.25)$$

For an ideal gas, we easily find

$$Z_{\mathrm{gr}}(\beta, V, \mu) = \exp\left[\mathrm{e}^{\beta\mu} V \left(\frac{2\pi m}{\beta h^2}\right)^{3/2}\right]. \quad (10.26)$$

The average internal energy and number of particles are

$$U = -\left(\frac{\partial \ln Z_{\mathrm{gr}}}{\partial \beta}\right)_{\beta\mu} = \frac{3}{2}\beta^{-1}\mathrm{e}^{\beta\mu} V \left(\frac{2\pi m}{\beta h^2}\right)^{3/2} \quad (10.27)$$

$$\bar{N} = \left(\frac{\partial \ln Z_{\mathrm{gr}}}{\partial(\beta\mu)}\right)_{\beta} = \mathrm{e}^{\beta\mu} V \left(\frac{2\pi m}{\beta h^2}\right)^{3/2} \quad (10.28)$$

from which we recover the relation $U = 3\bar{N}/2\beta$, now involving the average particle number.

10.4 Relation Between Statistical Mechanics and Thermodynamics

The highly successful science of thermodynamics deals with large systems in terms of macroscopic observable quantities alone. Equilibrium thermodynamics is derived, for the most part, from three basic principles, known as the zeroth, first and second laws, which summarize the phenomenological results of countless experiments. These principles are so well established by observation as to stand in no real need of further justification. However, our theoretical understanding would be seriously incomplete if we could not recover the results of thermodynamics from the microscopic laws of motion for the particles that constitute a macroscopic system. Moreover, once we can identify thermodynamic functions in statistical mechanical terms, we can set about obtaining predictions for their properties that cannot be obtained from thermodynamics alone. I am going to assume that readers are familiar with the principles of thermodynamics, but I shall first give a short summary of the points that particularly concern us. For simplicity, I shall deal explicitly only with fluid systems, but other systems, such as magnets and superconductors, which we shall need to consider later, can be dealt with by using straightforward analogies.

If two systems which are internally in equilibrium, their macroscopic properties having reached steady values, are brought into thermal contact, allowing heat energy to pass between them, their individual equilibria may be disturbed. If we wait long enough, however, the combined system will settle into a new equilibrium state, and we say that the two systems are in equilibrium with each other. The zeroth law of thermodynamics asserts that if two systems are simultaneously in equilibrium with a third, then they will be found to be in equilibrium with each other also. This implies that the systems share a common property, which has the same value for any two systems that are in equilibrium with each other. The property in question is *temperature*, and our discussion of the canonical ensemble indicates that β is a measure of thermodynamic temperature. The zeroth law does not, however, provide a means of assigning numerical values to temperature. Indeed, any property of a chosen standard system—a thermometer—that varies with temperature could be used to define an 'empirical scale of temperature'. Two such scales defined by different thermometers do not necessarily agree with each other.

The first law is essentially a statement of the conservation of energy, which explicitly recognizes that a change in the internal energy of a system can result equally from a flow of heat or from the performance of an equivalent amount of work. In a rudimentary way, we can distinguish a higher temperature from a lower one by agreeing, say, that the temperature of a system increases if heat flows into it and no work is done in the process.

The second law has been formulated in many different ways. The simplest, due in slightly different forms to Clausius and Kelvin, asserts that no process is possible whose only effect is the transfer of heat from a colder body to a hotter one. On the face of it, this is a purely qualitative statement, and it is quite remarkable that two precise, quantitative results follow from it. These are derived in every self-respecting textbook on thermodynamics. The first is that we can define an *absolute thermodynamic scale of temperature*. This scale is independent, in principle, of the properties of any specific system, but it coincides with the ideal gas scale, which is defined by the equation of state $pV = Nk_{\text{B}}T$ of an ideal gas (which holds for real gases in the limit that they become infinitely dilute). For an ideal gas, temperature is just a measure of the average kinetic energy of its molecules, and the value of k_{B} simply converts units of energy to the conventional units of temperature. The second result is that every equilibrium state of a system can be assigned an *entropy* S, in such a way that, if an amount of heat ΔQ flows into the system at a fixed temperature T, the change in entropy is $\Delta S = \Delta Q/T$. This actually defines the difference in entropy between any two equilibrium states, but not its absolute value.

Combining the first and second laws, we obtain the fundamental equation of the thermodynamics of fluids

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V \tag{10.29}$$

which expresses any change in internal energy as the sum of heat flow into the system and work done on it. In thermodynamic terms, this serves to define the pressure p. Because of this equation, the internal energy is naturally expressed as a function of the two quantities S and V, $U = U(S, V)$. This means that the partial derivatives $(\partial U/\partial S)_V$ and $(\partial U/\partial V)_S$ have recognizable physical interpretations as T and $-p$ respectively. While it is perfectly possible to write U as a function of, say, T and p, its partial derivatives with respect to these variables have no simple significance. If we wish to consider the possibility of particles entering or leaving the system, we extend (10.29) to read

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V + \mu\mathrm{d}N \tag{10.30}$$

where μ is the increase in internal energy due to the addition of a particle when no heat flow or performance of work accompanies the change. This provides the thermodynamic definition of the chemical potential.

The last two equations exemplify a general feature of thermodynamics, namely that a system can be characterized by a *thermodynamic potential.* This is a function of several macroscopic variables, which together specify the macroscopic state of the system, whose partial derivatives produce other quantities of physical interest. Several different functions may be used as potentials, and the criterion for a specific choice is that its natural independent variables should be quantities over which we exert experimental control. In statistical mechanics, we consider various idealized experimental situations in which systems are constrained in different ways and, as we have seen, these lead to different statistical ensembles. For a closed isoenergetic system, described by the microcanonical ensemble, the energy E (which for the moment I shall consider as identical to U), volume V and particle number N are all fixed and we need a potential for which these are the natural independent variables. By rearranging (10.30), we find

$$\mathrm{d}S = (1/T)\mathrm{d}E + (p/T)\mathrm{d}V - (\mu/T)\mathrm{d}N \tag{10.31}$$

which shows that the entropy $S(E, V, N)$ is a suitable choice.

For a closed, isothermal system, described by the canonical ensemble, the variables are T, V and N. The appropriate potential is the *Helmholtz free energy* $F = U - TS$. Using $\mathrm{d}(TS) = T\mathrm{d}S + S\mathrm{d}T$, we get

$$\mathrm{d}F = -S\mathrm{d}T - p\mathrm{d}V + \mu\mathrm{d}N \tag{10.32}$$

so indeed F is naturally expressed as $F(T, V, N)$. It is important to notice that we have done more than subtract TS from U. In (10.30), it is implied that both U and its partial derivatives T, p and μ are regarded as functions of S, V and N. In (10.32), it is similarly implied that F, S, p and μ are functions of T, V and N. This demands that we re-express S as a function of these variables by solving the equation

$$T = \left(\frac{\partial}{\partial S}U(S, V, N)\right)_{V,N} \tag{10.33}$$

for S. The whole process is a *Legendre transformation*, quite analogous to the passage from a Lagrangian to a Hamiltonian description of a classical dynamical system that we discussed in §3.3.

For an open isothermal system, described by the grand canonical ensemble, the independent variables are T, V and μ. By another Legendre transformation, we identify the appropriate potential as

$$\Omega(T, V, \mu) = F - \mu N = U - TS - \mu N \tag{10.34}$$

which is called the *grand potential*. The following argument allows us to relate the grand potential more directly to observable macroscopic quantities. We return to the entropy $S(E, V, N)$ and observe that all four of the variables S, E, V and N are *extensive*. That is to say, they are all proportional to the total size of the system. If we increase the total size by a factor λ, so that it contains λN particles in a volume λV and these particles have a total energy λE, then we see intuitively that any small part of the enlarged system should look exactly the same as a similar small part of the original system. (This intuition might fail us in some circumstances. If, for example, there are interparticle forces whose range is comparable with the size of the whole system, then the state of some small part might depend on the total size. Here, I am ignoring such possibilities.) An amount of heat ΔQ flowing into the original system should have the same effect on any small part as an amount $\lambda\Delta Q$ flowing into the enlarged system, so the enlarged system has entropy λS. The entropy must therefore be a *homogeneous function*, in the sense that

$$S(\lambda E, \lambda V, \lambda N) = \lambda S(E, V, N). \tag{10.35}$$

Let us differentiate this equation with respect to λ and then set $\lambda = 1$. We find

$$E\frac{\partial S}{\partial E} + V\frac{\partial S}{\partial V} + N\frac{\partial S}{\partial N} = S. \tag{10.36}$$

The various partial derivatives can be identified from (10.31) and we discover the relation (still taking E to be equivalent to U)

$$TS = U + pV - \mu N \tag{10.37}$$

which implies that $\Omega(T, V, \mu) = -Vp(T, V, \mu)$. Readers should not find it hard to see that p, T and μ are *intensive* variables, being independent of the total size of the system, and that p therefore cannot depend on V independently of T and μ. (For example, in the ideal-gas equation of state, $p = (N/V)k_{\mathrm{B}}T$, the values of T and μ determine the number of particles per unit volume N/V, as can be seen from (10.28).) We can thus write the grand potential as

$$\Omega(T, V, \mu) = -Vp(T, \mu). \tag{10.38}$$

These three potentials can be identified in terms of the statistical partition functions $\Sigma(E, V, N)$, $Z_{\mathrm{can}}(\beta, V, N)$ and $Z_{\mathrm{gr}}(\beta, V, \mu)$. To do this safely, however, it is necessary to consider the *thermodynamic limit* in which N and V are taken to infinity, with the number of particles per unit volume N/V held fixed. The reason for this is that, in thermodynamics, it is assumed that the quantities U, T, μ and N all have definite values. In statistical mechanics this is not true. In an isothermal system, for example, the temperature is fixed by an infinite heat bath, but the energy fluctuates and U can be identified only as an average energy. In an isoenergetic system, by contrast, the energy E is fixed. Because the interpretation of the variables varies from one ensemble to another, the entropy, Helmholtz free energy and grand potential obtained from the appropriate ensembles will not be related by the thermodynamic Legendre transformations unless the effect of fluctuations is negligible. I said earlier that relative fluctuations are expected to be proportional to $N^{-1/2}$, and readers are encouraged to investigate this in exercise 10.3. If so, then we can expect to obtain a unique correspondence between statistical mechanics and thermodynamics in the thermodynamic limit. Experimentally, we deal with systems of finite size, but typical numbers of particles are of the order of Avogadro's number 6.02×10^{23} which is, to a fair approximation, infinite!

Let us start with the grand canonical ensemble and define

$$\Omega_{\mathrm{gr}}(T, V, \mu) = -k_{\mathrm{B}}T \ln Z_{\mathrm{gr}}(\beta, V, \mu). \tag{10.39}$$

We would like to identify this as the grand canonical version of the thermodynamic potential $\Omega(T, V, \mu)$. If we can identify its partial derivatives with respect to T, V and μ as $-S$, $-p$ and $-\beta N$ respectively, then the two functions can differ only by an additive constant, which can be determined by direct calculation if necessary. It follows from (10.28) that the μ derivative is $-\beta\bar{N}$, and in the thermodynamic limit we identify the mean number of particles $\bar{N}$ with the thermodynamic variable N. For the T derivative, we can use (10.17), (10.27) and (10.28) to find

$$\begin{aligned}\frac{\partial\Omega_{\mathrm{gr}}}{\partial T} = -\frac{1}{k_{\mathrm{B}}T^2}\frac{\partial\Omega_{\mathrm{gr}}}{\partial\beta} &= \frac{1}{T}\left[\Omega_{\mathrm{gr}} + \left(\frac{\partial \ln Z_{\mathrm{gr}}}{\partial\beta}\right)_{\beta\mu} + \mu\left(\frac{\partial \ln Z_{\mathrm{gr}}}{\partial(\beta\mu)}\right)_{\beta}\right]\\ &= \frac{1}{T}(\Omega_{\mathrm{gr}} - U + \mu\bar{N}).\end{aligned} \tag{10.40}$$

We do not have a definition of entropy within the grand canonical ensemble, but we can argue self-consistently that if, indeed, (10.39) is the correct grand-canonical version of Ω then, according to (10.34), the appropriate definition must be $S_{\mathrm{gr}} = -T^{-1}(\Omega_{\mathrm{gr}} - U + \mu\bar{N})$, in which case we have shown that $\partial\Omega_{\mathrm{gr}}/\partial T = -S_{\mathrm{gr}}$, as required. Similarly, we have no grand-canonical definition of the pressure, so we must resort to defining $p_{\mathrm{gr}} = -\partial\Omega_{\mathrm{gr}}/\partial V$. We can check that this is, at least, sensible in the case of an ideal gas, by using (10.26) and (10.28) to recover the equation of state $p_{\mathrm{gr}} = \bar{N}k_{\mathrm{B}}T/V$.

Readers may like to develop similar arguments to show that, for the canonical ensemble,

$$F_{\mathrm{can}}(T, V, N) = -k_{\mathrm{B}}T \ln Z_{\mathrm{can}}(\beta, V, N) \tag{10.41}$$

and for the microcanonical ensemble

$$S_{\mathrm{mico}}(E, V, N) = k_{\mathrm{B}} \ln\left(\frac{\Sigma(E, V, N)}{h^{3N}N!}\right). \tag{10.42}$$

I shall follow the alternative course of showing that, in the thermodynamic limit, these functions are obtained from (10.39) by the thermodynamic Legendre transformations. Consider equation (10.25). In the thermodynamic limit, we expect fluctuations in N to be small relative to $\bar{N}$, so only those terms in the sum for which $N \approx \bar{N}$ should make significant contributions. We can therefore make the estimate

$$Z_{\text{gr}}(\beta, V, \mu) = K\mathrm{e}^{\beta\mu\bar{N}} Z_{\text{can}}(\beta, V, \bar{N}) \tag{10.43}$$

where K represents the number of important terms. We now use (10.39) and (10.41) to write

$$\frac{\Omega_{\text{gr}}}{\bar{N}} = \frac{F_{\text{can}}}{\bar{N}} - \mu - \frac{\ln K}{\bar{N}}. \tag{10.44}$$

In the thermodynamic limit, we expect the potentials to be extensive, in the sense I explained earlier on. The quantity K is not precisely defined, but it should depend only weakly on N. In the thermodynamic limit, therefore, the last term in (10.44) vanishes and the remaining equation coincides with (10.34). Both potentials can now be obtained from either ensemble with the same result, so we have a unique correspondence with thermodynamics and the ensemble subscripts can be dropped.

A relation between the canonical and microcanonical ensembles can be derived in a similar manner. Using (10.7) and (10.14), we can write

$$\begin{aligned} Z_{\text{can}}(\beta, V, N) &= \frac{1}{h^{3N}N!}\int \mathrm{d}E \int \mathrm{d}^{6N}X\, \mathrm{e}^{-\beta E}\delta[E - H_N(X)] \\ &= \frac{1}{h^{3N}N!}\int \mathrm{d}E\, \mathrm{e}^{-\beta E}\Sigma(E, V, N). \end{aligned} \tag{10.45}$$

Then, treating fluctuations in energy in the same way as those in the number of particles, and using the definitions (10.41) and (10.42) of the canonical and microcanonical potentials, we recover the thermodynamic relation $F = U - TS$. In this way, we see that all three statistical ensembles become equivalent in the thermodynamic limit and their partition functions can be uniquely identified in terms of thermodynamic potentials. Mathematically, it is interesting to note that the Legendre transforms which relate these potentials correspond to *Laplace transforms* which relate the partition functions. The arguments I used to derive these relations are, of course, by no means rigorous. In principle, assumptions such as the extensivity of the potentials should be checked for each system to which the theory is applied. Indeed, it is possible to invent theoretical models for which the arguments do not work. For example, as suggested by earlier remarks, the thermodynamic limit may not exist when there are long-range forces. As far as I know, the arguments are sound for all systems of physical interest. Readers may like to check for themselves that everything goes through smoothly for the ideal gas. They should find that the entropy is given by the *Sackur–Tetrode equation*

$$\frac{S}{N} = k_{\mathrm{B}} \left\{ \frac{5}{2} + \ln \left[\frac{V}{N} \left(\frac{2\pi m k_{\mathrm{B}} T}{h^2} \right)^{3/2} \right] \right\}. \tag{10.46}$$

Factors of $N!$ should be treated using Stirling's approximation

$$\ln(N!) = N \ln(N) - N + \tfrac{1}{2} \ln(2\pi N) + \ldots \tag{10.47}$$

valid for large N.

10.5 Quantum Statistical Mechanics

When dealing with a large quantum-mechanical system, we need to estimate the expectation values of operators in states that we are unable to specify exactly at a microscopic level. We therefore have to take two averages, one over the uncertainties inherent in a definite quantum state and one to take account of our ignorance of what the state actually is. For the time being, I shall work in the Schrödinger picture. Suppose we have a complete orthonormal set of states $|\psi_n(t)\rangle$ for which

$$\langle \psi_m(t) | \psi_n(t) \rangle = \delta_{mn} \qquad \text{and} \qquad \sum_n |\psi_n(t)\rangle \langle \psi_n(t)| = \hat{I}. \tag{10.48}$$

For simplicity, I am assuming that these states can be labelled by a discrete index n; there will be no difficulty in converting the sums into integrals where necessary. Suppose further that we can specify for each state the probability P_n of finding the system in that state. As long as the system is left undisturbed, P_n does not change with time. Using (10.48), we can write the expectation value of an observable A at time t as

$$\bar{A}(t) = \sum_n \langle \psi_n(t) | \hat{A} | \psi_n(t) \rangle P_n = \sum_{m,n} \langle \psi_m(t) | \hat{A} | \psi_n(t) \rangle P_n \langle \psi_n(t) | \psi_m(t) \rangle. \tag{10.49}$$

The object

$$\hat{\rho}(t) = \sum_n |\psi_n(t)\rangle P_n \langle \psi_n(t)| \tag{10.50}$$

can be regarded as an operator, called the *density operator*, which acts on a bra or ket vector to produce another:

$$\langle \Psi | \hat{\rho} = \sum_n \left[\langle \Psi | \psi_n(t) \rangle P_n \right] \langle \psi_n(t)| \quad \text{or} \quad \hat{\rho} |\Psi\rangle = \sum_n |\psi_n(t)\rangle \left[P_n \langle \psi_n(t) | \Psi \rangle \right]. \tag{10.51}$$

The expectation value (10.49) is the sum of diagonal matrix elements of $\hat{A}\,\hat{\rho}$, which is the trace of $\hat{A}\,\hat{\rho}$:

$$\bar{A}(t) = \sum_m \langle \psi_m(t) | \hat{A}\,\hat{\rho} | \psi_m(t) \rangle = \mathrm{Tr}[\hat{A}\,\hat{\rho}\,]. \tag{10.52}$$

It is readily verified that

$$\mathrm{Tr}[\hat{\rho}\,\hat{A}] = \mathrm{Tr}[\hat{A}\,\hat{\rho}\,] \tag{10.53}$$

and, on account of the normalization of probabilities, that

$$\mathrm{Tr}[\hat{\rho}\,] = \sum_n P_n = 1. \tag{10.54}$$

The density operator behaves rather differently from the operators that represent observable quantities. Because it is constructed from state vectors that represent possible histories of the system, it is time-dependent in the Schrödinger picture and time-independent in the Heisenberg picture. In the Schrödinger picture, we can use the Schrödinger equation (5.32) with (5.33) to obtain the equation of motion

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}(t) = \frac{\mathrm{i}}{\hbar}[\hat{\rho}(t), \hat{H}] \tag{10.55}$$

which is the quantum-mechanical version of the Liouville equation (10.4). It differs by a minus sign from the equation of motion (5.36) for time-dependent operators that represent observables in the Heisenberg picture.

The arguments we used to derive the ensembles of classical statistical mechanics can be taken over to the quantum theory. To describe thermal equilibrium, we want the density operator to be time independent in the Schrödinger picture. According to (10.55), it must therefore be constructed from operators which commute with the Hamiltonian, including the Hamiltonian itself. For a system of N particles confined to a volume V, we obtain the canonical density operator as

$$\hat{\rho}_{\mathrm{can}} = Z_{\mathrm{can}}^{-1}\exp(-\beta\hat{H}_N) \tag{10.56}$$

where the partition function is given by

$$Z_{\mathrm{can}}(\beta, V, N) = \mathrm{Tr}\left[\exp(-\beta\hat{H}_N)\right]. \tag{10.57}$$

No factor of h^{-3N} is required because this expression is already dimensionless, and no factor of $1/N!$, because the indistinguishability of identical particles is taken into account in the definition of the quantum states. The grand partition function may be defined by analogy with (10.25) as

$$Z_{\mathrm{gr}}(\beta, V, \mu) = \sum_N \exp(\beta\mu N) Z_{\mathrm{can}}(\beta, V, N). \tag{10.58}$$

Alternatively, we can resort to second quantization and define the grand-canonical density operator and partition function by

$$\hat{\rho}_{\mathrm{gr}} = Z_{\mathrm{gr}}^{-1}\exp[-\beta(\hat{H} - \mu\hat{N})] \tag{10.59}$$

$$Z_{\mathrm{gr}}(\beta, V, \mu) = \mathrm{Tr}\left\{\exp[-\beta(\hat{H} - \mu\hat{N})]\right\}. \tag{10.60}$$

Here, of course, the trace includes states with any number of particles. When the number of particles is not conserved, it makes no sense to speak of a fixed number. Moreover, the particle number operator $\hat{N}$ does not commute with the Hamiltonian (in fact, it may not even be well defined) and cannot appear in the equilibrium density operator. In that case, we must use (10.59) and (10.60) with $\mu = 0$. It is a matter of taste whether this is regarded as a grand-canonical description of a system of particles or, on the other hand, as a canonical description of the underlying system of quantum fields.

Quantum-mechanical ideal gases are most conveniently treated in the grand canonical ensemble. Since the particles do not interact, eigenstates of the operator $\hat{H} - \mu\hat{N}$ can be built from single-particle energy eigenstates. If we consider a gas confined to a cubical box of side L, the single-particle momentum eigenstates have momenta

$$\boldsymbol{p} = (h/L)\boldsymbol{i} \tag{10.61}$$

where $\boldsymbol{i}$ is a triplet of integers, each of which can have any positive or negative value. If the particles have spin s, then for each momentum value, with single-particle energy $\epsilon_{\boldsymbol{i}} = p_{\boldsymbol{i}}^2/2m$, there are $(2s+1)$ independent spin polarization states. We now take the states $|\psi_n\rangle$ to be the basis states of the occupation number representation, with $n_{\boldsymbol{i}\sigma}$ particles in the state with momentum labelled by $\boldsymbol{i}$ and spin polarization σ. The grand partition function is

$$Z_{\text{gr}} = \sum_{\{n_{\boldsymbol{i}\sigma}\}} \exp\left[-\beta\sum_{\boldsymbol{i},\sigma}(\epsilon_{\boldsymbol{i}} - \mu)n_{\boldsymbol{i}\sigma}\right] = \prod_{\boldsymbol{i},\sigma}\sum_{\{n_{\boldsymbol{i}\sigma}\}} \exp[-\beta(\epsilon_{\boldsymbol{i}} - \mu)n_{\boldsymbol{i}\sigma}]. \tag{10.62}$$

For bosons, each $n_{\boldsymbol{i}\sigma}$ ranges from 0 to ∞, while for fermions it takes only the values 0 or 1. In either case, all the sums can be carried out (for bosons, the infinite sum is a geometric series) giving

$$Z_{\text{gr}} = \prod_{\boldsymbol{i}} \{1 \pm \exp[-\beta(\epsilon_{\boldsymbol{i}} - \mu)]\}^{\pm(2s+1)} \tag{10.63}$$

where the upper signs refer to fermions and the lower ones to bosons. The average occupation numbers of single-particle momentum states are easily found:

$$\bar{n}_{\boldsymbol{i}} = \sum_{\sigma} \bar{n}_{\boldsymbol{i}\sigma} = -\frac{\partial \ln Z_{\text{gr}}}{\partial(\beta\epsilon_{\boldsymbol{i}})} = (2s+1)\left\{\exp[\beta(\epsilon_{\boldsymbol{i}} - \mu)] \pm 1\right\}^{-1}. \tag{10.64}$$

Under all circumstances of practical interest, sums over momentum states can be replaced with integrals, and (10.61) leads to the replacement $\sum_{\boldsymbol{i}} \to (V/h^3)\int \mathrm{d}^3p$, where $V = L^3$ is the volume. The energy becomes $\epsilon = \boldsymbol{p}^2/2m$ and, since this depends only on the magnitude of $\boldsymbol{p}$, the angular integrals over the direction of $\boldsymbol{p}$ can be carried out. After defining $x = (\beta/2m)^{1/2}|\boldsymbol{p}|$, we find for the logarithm of the partition function

$$\ln Z_{\text{gr}} = \pm 4\pi V(2s+1)\left(\frac{2m}{\beta h^2}\right)^{3/2}\int_0^\infty \mathrm{d}x\, x^2 \ln(1 \pm z\mathrm{e}^{-x^2}) \tag{10.65}$$

and for the average number of particles per unit volume

$$\frac{\bar{N}}{V} = 4\pi(2s+1)\left(\frac{2m}{\beta h^2}\right)^{3/2} z\int_0^\infty \mathrm{d}x\, x^2 \mathrm{e}^{-x^2}(1 \pm z\mathrm{e}^{-x^2})^{-1} \tag{10.66}$$

where z is the fugacity (10.24). At low temperatures, quantum ideal gases behave very differently from classical ones. I shall discuss some of the low-temperature properties of bosons in the next chapter. The case of fermions, which I shall not discuss, is particularly important when applied to the gas of electrons in a metal and is dealt with extensively in most textbooks on solid state physics. At high temperatures, on the other hand, quantum gases differ very little from classical ones. From (10.66), we see that if β becomes very small with $\bar{N}/V$ fixed, then the fugacity z must also become small. In that case, (10.65) can be approximated as

$$\ln Z_{\mathrm{gr}} \approx 4\pi V(2s+1)\left(\frac{2m}{\beta h^2}\right)^{3/2} z\int_0^\infty \mathrm{d}x\, x^2\mathrm{e}^{-x^2} = (2s+1)zV\left(\frac{2\pi m}{\beta h^2}\right)^{3/2}. \tag{10.67}$$

This agrees exactly with (10.26), apart from the spin multiplicity factor $(2s+1)$. For spin-0 particles, which can be compared most directly with their classical counterparts, this factor is 1. For particles with higher spin, the familiar relations $U = 3Nk_{\mathrm{B}}T/2$ and $pV = Nk_{\mathrm{B}}T$ are unaffected.

10.6 Field Theories at Finite Temperature

Although we have found it possible to treat ideal gases without any detailed use of second quantization, field-theoretic methods are more or less essential for the systematic study of large systems of interacting particles. We have seen, moreover, that relativistic particles can be correctly described only by a quantum field theory. It is therefore necessary to find methods of evaluating quantities such as (10.52) or (10.60) when $\hat{H}$ and $\hat{N}$ are second-quantized operators. A useful technique comes about from realizing that each of the matrix elements in the trace in (10.60) is analogous to the one we evaluated in (9.28), if we replace $\hat{H}$ with $\hat{H} - \mu\hat{N}$ and $t_{\mathrm{f}} - t_{\mathrm{i}}$ with $-\mathrm{i}\beta$. This leads to the *imaginary-time formalism*, in which the diagrammatic perturbation theory we discussed in chapter 9 can be taken over more or less intact, simply by replacing real time t with an imaginary time $\tau = \mathrm{i}t$. This imaginary time takes values between 0 and β. Here, I shall discuss only the case of a relativistic scalar field ϕ, but other relativistic and non-relativistic field theories can be treated by similar methods.

Since we are considering a many-particle system in thermal equilibrium, its rest frame is a preferred frame of reference. Therefore, even in a relativistic theory, there is a preferred measure of time, namely that measured in the rest frame, which provides a natural means of distinguishing Heisenberg and Schrödinger pictures. For simplicity, I shall take the chemical potential to be zero.

If $\hat{\phi}(\boldsymbol{x})$ is the Schrödinger-picture field operator, then we define the *imaginary-time Heisenberg picture* by

$$\hat{\phi}(\boldsymbol{x},\tau) = \mathrm{e}^{\hat{H}\tau}\hat{\phi}(\boldsymbol{x})\mathrm{e}^{-\hat{H}\tau} \quad \text{and} \quad \hat{\phi}^\dagger(\boldsymbol{x},\tau) = \mathrm{e}^{\hat{H}\tau}\hat{\phi}^\dagger(\boldsymbol{x})\mathrm{e}^{-\hat{H}\tau}. \tag{10.68}$$

It should be noticed that $\hat{\phi}^\dagger(\boldsymbol{x},\tau)$ is *not* the adjoint of $\hat{\phi}(\boldsymbol{x},\tau)$ in the usual sense. By analogy with (9.14), we define an imaginary-time propagator by

$$G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau') = \mathrm{Tr}\left[\hat{\rho}\,T_\tau[\hat{\phi}(\boldsymbol{x},\tau)\hat{\phi}^\dagger(\boldsymbol{x}',\tau')]\right] \tag{10.69}$$

where T_τ is the latest-on-the-left ordering operator for imaginary times. This propagator will indeed depend only on $\boldsymbol{x}-\boldsymbol{x}'$ if the equilibrium state is homogeneous, as intuitively it must be. By using the identity $\mathrm{Tr}(\hat{A}\hat{B}) = \mathrm{Tr}(\hat{B}\hat{A})$, valid for any $\hat{A}$ and $\hat{B}$, it is easy to show that it also depends only on $\tau-\tau'$.

The same identity may be used to derive a vital property of the propagator, namely that it is periodic in $\tau-\tau'$, with period β. That is

$$G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'+\beta) = G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'). \tag{10.70}$$

Since τ and τ' both lie between 0 and β, their difference lies between $-\beta$ and β, so (10.70) is meaningful only when $\tau < \tau'$. On the other hand, $\tau+\beta$ must be greater than τ', so we have

$$\begin{aligned} G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'+\beta) &\\ &= Z_{\mathrm{gr}}^{-1}\,\mathrm{Tr}\left[\mathrm{e}^{-\beta\hat{H}}\mathrm{e}^{(\tau+\beta)\hat{H}}\hat{\phi}(\boldsymbol{x})\mathrm{e}^{-(\tau+\beta)\hat{H}}\mathrm{e}^{\tau'\hat{H}}\hat{\phi}^\dagger(\boldsymbol{x}')\mathrm{e}^{-\tau'\hat{H}}\right] \\ &= Z_{\mathrm{gr}}^{-1}\,\mathrm{Tr}\left[\mathrm{e}^{\tau\hat{H}}\hat{\phi}(\boldsymbol{x})\mathrm{e}^{-\tau\hat{H}}\mathrm{e}^{-\beta\hat{H}}\mathrm{e}^{\tau'\hat{H}}\hat{\phi}^\dagger(\boldsymbol{x}')\mathrm{e}^{-\tau'\hat{H}}\right] \\ &= Z_{\mathrm{gr}}^{-1}\,\mathrm{Tr}\left[\mathrm{e}^{-\beta\hat{H}}\mathrm{e}^{\tau'\hat{H}}\hat{\phi}^\dagger(\boldsymbol{x}')\mathrm{e}^{-\tau'\hat{H}}\mathrm{e}^{\tau\hat{H}}\hat{\phi}(\boldsymbol{x})\mathrm{e}^{-\tau\hat{H}}\right]. \end{aligned} \tag{10.71}$$

For $\tau < \tau'$, this is indeed equal to $G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau')$. For $\tau > \tau'$, the corresponding relation

$$G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'-\beta) = G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau') \tag{10.72}$$

can be established in the same way. In the case of fermions, the propagator is antiperiodic, which means that

$$S(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'\pm\beta) = -S(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'). \tag{10.73}$$

The expectation value of any operator constructed from the fields can, in principle, be calculated from the propagator or from other imaginary-time Green functions. For example, to obtain the expectation value of $\hat{\phi}^\dagger(\boldsymbol{x})\hat{\phi}(\boldsymbol{x})$, we would use

$$\begin{aligned} \langle\hat{\phi}^\dagger(\boldsymbol{x})\hat{\phi}(\boldsymbol{x})\rangle = \mathrm{Tr}\left[\hat{\rho}\,\hat{\phi}^\dagger(\boldsymbol{x})\hat{\phi}(\boldsymbol{x})\right] &= \lim_{\epsilon\to 0}\mathrm{Tr}\left[\hat{\rho}\,T_\tau[\hat{\phi}(\boldsymbol{x},\tau)\hat{\phi}^\dagger(\boldsymbol{x},\tau+\epsilon)]\right] \\ &= \lim_{\epsilon\to 0} G(\boldsymbol{0},-\epsilon). \end{aligned} \tag{10.74}$$

The Green functions in turn can be represented by functional integrals similar to (9.32), except that these must be converted to imaginary time. The result, derived by a method similar to that of §9.3, is

$$\begin{aligned}\mathrm{Tr}\left[\hat{\rho}\, T_\tau[\hat{\phi}(x_1)\cdots\ \hat{\phi}^\dagger(x_n)]\right]&\\ &= Z_{\mathrm{gr}}^{-1}\int \mathcal{D}\phi(x)\,\phi(x_1)\cdots\phi^*(x_n)\exp[-S_\beta(\phi)]\end{aligned} \tag{10.75}$$

where $\phi(x)$ means $\phi(\boldsymbol{x},\tau)$ and the symbol $\mathcal{D}\phi(x)$ includes a normalizing factor to make (10.54) true. The finite-temperature action S_β is found by replacing t with $-\mathrm{i}\tau$. For the self-interacting scalar field we studied in chapter 9, it is given by

$$S_\beta(\phi) = \int_0^\beta \mathrm{d}\tau\int \mathrm{d}^3x\left[\frac{\partial\phi^*}{\partial\tau}\frac{\partial\phi}{\partial\tau} + \nabla\phi^*\cdot\nabla\phi + m^2\phi^*\phi + \frac{\lambda}{4}(\phi^*\phi)^2\right]. \tag{10.76}$$

Proceeding as in chapter 9, we can find the equation analogous to (9.37) satisfied by the unperturbed propagator $G_0(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau')$, namely

$$\left(\frac{\partial^2}{\partial\tau^2} + \nabla^2 - m^2\right)G_0(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau') = -\delta(\tau-\tau')\delta(\boldsymbol{x}-\boldsymbol{x}'). \tag{10.77}$$

Because of the periodicity in imaginary time, we express this propagator in terms of a Fourier transform as

$$\begin{aligned}&G_0(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau')\\ &\quad= \int\frac{\mathrm{d}^3k}{(2\pi)^3}\exp[\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')]\beta^{-1}\sum_{n=-\infty}^{\infty}\exp[\mathrm{i}\omega_n(\tau-\tau')]\widetilde{G}_0(\boldsymbol{k},n)\end{aligned} \tag{10.78}$$

where $\omega_n = 2\pi n/\beta$. The frequencies ω_n are known as *Matsubara frequencies*. On substituting in (10.77), we find

$$\widetilde{G}_0(\boldsymbol{k},n) = \left(k^2 + \omega_n^2 + m^2\right)^{-1}. \tag{10.79}$$

To see how the finite-temperature field theory fits in with our earlier discussion of quantum gases, let us evaluate $\ln Z_{\mathrm{gr}}$ for the case of an ideal relativistic gas, with $\lambda = 0$. The partition function provides the normalizing factor in (10.75), and since $\mathrm{Tr}[\hat{\rho}] = 1$, it is clearly given by

$$Z_{\mathrm{gr}} = \int\mathcal{D}\phi(x)\ \exp[-S_\beta(\phi)]. \tag{10.80}$$

This, however, is slightly ambiguous because of an ill-defined constant that appears in the definition of the functional integral (see (9.28), for example). To avoid this difficulty, we can calculate the quantity $-\partial \ln Z_{\text{gr}}/\partial m^2$ which, as we see from (10.76), is given by

$$-\frac{\partial \ln Z_{\text{gr}}}{\partial m^2} = \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \, \langle \phi^*(\boldsymbol{x}, \tau)\phi(\boldsymbol{x}, \tau)\rangle = \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \, G_0(\boldsymbol{0}, 0). \tag{10.81}$$

Since $G_0(\boldsymbol{0}, 0)$ is independent of $\boldsymbol{x}$ and τ, the two integrals just give a factor of βV. To evaluate $G_0(\boldsymbol{0}, 0)$, we use the identity

$$\sum_{n=-\infty}^{\infty} \frac{1}{n^2 + a^2} = \frac{\pi}{a}\left(\frac{\mathrm{e}^{\pi a} + \mathrm{e}^{-\pi a}}{\mathrm{e}^{\pi a} - \mathrm{e}^{-\pi a}}\right) = \frac{\pi}{a}\coth(\pi a) \tag{10.82}$$

which readers are invited to prove in exercise 10.7. We obtain

$$\begin{aligned} -\frac{\partial \ln Z_{\text{gr}}}{\partial m^2} &= V \int \frac{\mathrm{d}^3k}{(2\pi)^3} \frac{\beta}{2\omega(\boldsymbol{k})} \coth(\tfrac{1}{2}\beta\omega(\boldsymbol{k})) \\ &= \frac{\partial}{\partial m^2}\left[2V \int \frac{\mathrm{d}^3k}{(2\pi)^3} \ln\left(\mathrm{e}^{\beta\omega(\boldsymbol{k})/2} - \mathrm{e}^{-\beta\omega(\boldsymbol{k})/2}\right)\right]. \end{aligned} \tag{10.83}$$

Up to a possible constant of integration, this gives

$$\begin{aligned} -\ln Z_{\text{gr}} = 2V \frac{1}{2\pi^2\beta^3} \int_0^\infty \mathrm{d}x \, x^2 \ln\left[1 - \exp[-(x^2 + \beta^2 m^2)^{1/2}]\right] \\ + \beta V \int \frac{\mathrm{d}^3k}{(2\pi)^3} \omega(\boldsymbol{k}). \end{aligned} \tag{10.84}$$

Remembering that the internal energy is $U = -\partial \ln Z_{\text{gr}}/\partial\beta$, we recognize the last term as the infinite vacuum energy encountered in (7.21), as long as we identify $(2\pi)^3\delta(0) = \int \mathrm{d}^3x = V$. The first term, in which $x = \beta|\boldsymbol{k}|$, is obviously similar to (10.65) with $z = 1$. The field theory describes particles of spin $s = 0$, and the overall factor of 2 represents the two equal contributions from particles and antiparticles. Other differences arise from the relativistic energy relation $\omega(\boldsymbol{k}) = (\boldsymbol{k}^2 + m^2)^{1/2}$ and the use of natural units, in which $h = 2\pi$. The non-relativistic limit of (10.84) is explored in exercise 10.8.

10.7 Black Body Radiation

Black-body radiation is most simply conceived of as an ideal gas of photons in thermal equilibrium with the walls of a cavity that contains it. According to quantum electrodynamics, photons can scatter from each other by way of intermediate states containing virtual charged particles. Under almost all circumstances, however, this interaction is entirely negligible. Because photons are massless, there is no lower limit to the energy change involved in the emission or absorption of a photon by the cavity walls. There is therefore no constraint on the total number of photons in the gas and its chemical potential is zero. It is possible to derive the partition function from QED, but problems are again encountered with redundant gauge degrees of freedom. In particular, the treatment of the component A_0 of the vector potential in the imaginary-time formalism needs careful consideration. I shall not discuss these questions in detail. It should come as no surprise, though, that we obtain the correct result simply by setting $m = 0$ in (10.84). Since photons are their own antiparticles, the overall factor of 2 arises in this case from the two independent spin polarization states.

At very high temperatures, such as we shall later encounter in connection with the early universe, a modified version of black-body radiation arises, in which any particle species whose mass is much smaller than $k_{\mathrm{B}}T$ can be considered effectively massless and treated on the same footing as photons. As long as an ideal-gas description remains appropriate, we simply add the contributions to $\ln Z_{\mathrm{gr}}$ from each species. If we drop the unobservable vacuum energy, then, for each bosonic species, the contribution is

$$-\ln Z_{\mathrm{gr}} = g_{\mathrm{b}} V \frac{1}{2\pi^2\beta^3} I_{\mathrm{b}} \tag{10.85}$$

where g_{b} is the number of independent spin polarization states of particles and antiparticles and

$$I_{\mathrm{b}} = \int_0^\infty \mathrm{d}x\, x^2 \ln(1 - \mathrm{e}^{-x}) = -\frac{\pi^4}{45}. \tag{10.86}$$

For fermions, this integral is modified in the same way as that in (10.65). It is given by

$$I_{\mathrm{f}} = -\int_0^\infty \mathrm{d}x\, x^2 \ln(1 + \mathrm{e}^{-x}) = \tfrac{7}{8} I_{\mathrm{b}}. \tag{10.87}$$

In view of this relation (which is readily verified by showing that $I_{\mathrm{b}} - I_{\mathrm{f}} = I_{\mathrm{b}}/8$), we can treat the gas as a whole by defining

$$g = \sum_{\text{boson species}} g_{\mathrm{b}} + \tfrac{7}{8} \sum_{\text{fermion species}} g_{\mathrm{f}}. \tag{10.88}$$

To return to laboratory units, we must divide $\ln Z_{\mathrm{gr}}$ by $(\hbar c)^3$ to make it dimensionless. We then have

$$\Omega = -k_{\mathrm{B}}T \ln Z_{\mathrm{gr}} = -V \frac{2g\sigma}{3c} T^4 \tag{10.89}$$

where

$$\sigma = \frac{\pi^2 k_{\mathrm{B}}{}^4}{60\hbar^3 c^2} = 5.6698 \times 10^{-8}\,\mathrm{W\,m^{-2}K^{-4}} \tag{10.90}$$

is the Stefan–Boltzmann constant. It is a simple matter to derive the following expressions for the energy and entropy densities and the pressure:

$$\frac{U}{V} = -\frac{1}{V}\frac{\partial \ln Z_{\mathrm{gr}}}{\partial \beta} = \frac{2g\sigma}{c}T^4 \tag{10.91}$$

$$\frac{S}{V} = -\frac{1}{V}\frac{\partial \Omega}{\partial T} = \frac{8g\sigma}{3c}T^3 \tag{10.92}$$

$$p = -\frac{\partial \Omega}{\partial V} = \frac{2g\sigma}{3c}T^4 = \frac{1}{3}\frac{U}{V}. \tag{10.93}$$

10.8 The Classical Lattice Gas

Our explicit examples have so far been restricted to ideal gases, because the approximation methods needed to treat non-ideal gases and liquids require quite lengthy development, for which there is no space in this book. I shall, however, describe a straightforward, if somewhat crude, approximation to a non-ideal classical gas, which is of some importance in the theory of phase transitions. This is the *lattice gas*. We consider a gas whose molecules interact through a pair potential $W(r)$, so the Hamiltonian for N molecules is

$$H_N = \sum_{i=1}^{N} \frac{1}{2m}\boldsymbol{p}_i^2 + \frac{1}{2}\sum_{i,j=1}^{N} W(|\boldsymbol{x}_i - \boldsymbol{x}_j|). \tag{10.94}$$

Inserting this into (10.14), we find that the momentum integrals can be carried out, so the canonical partition function is

$$Z_{\mathrm{can}}(\beta, V, N) = \left(\frac{2\pi m}{\beta h^2}\right)^{3N/2} \frac{1}{N!}\int \mathrm{d}^{3N}x \, \exp\left(-\tfrac{1}{2}\beta \sum_{i,j=1}^{N} W(|\boldsymbol{x}_i - \boldsymbol{x}_j|)\right). \tag{10.95}$$

The remaining integral is a sum over all instantaneous configurations of the positions of the molecules, and there is some advantage to re-expressing this sum in the following approximate manner. Real molecules exhibit a strong repulsion at short distances, so it makes sense to divide the total volume occupied by the gas into a large number of cells, each having a volume v comparable with the volume of a single molecule, and to suppose that there can be at most one molecule in any one cell. The mid-points of the cells will usually be taken to form a regular lattice in space. To the ith cell, we assign an occupation number n_i, which is 1 for an occupied cell or 0 for an empty one; the sum of the n_i for all the cells is the total number of molecules N. We now take the potential energy of a pair of molecules to depend only upon the cells occupied by the molecules, but not on their precise location within the cells, which will be a reasonable approximation for potentials that vary little over the size of a cell. For a given set of N occupied cells, the integral in (10.95) now gives v^N for each of the $N!$ distributions of N

molecules in the N cells. By summing over all possible sets of N occupied cells, we obtain

$$Z_{\text{can}}(\beta, V, N) = \left(\frac{2\pi m}{\beta h^2}\right)^{3N/2} v^N \sum_{\{n_i\}}^{(N)} \exp\left(-\tfrac{1}{2}\beta \sum_{i,j} W_{ij} n_i n_j\right) \quad (10.96)$$

where i and j now label all the cells in the lattice and W_{ij} is the potential between particles in cells i and j when these cells are occupied. The configuration sum is over all sets of values $n_i = 0, 1$ consistent with their sum being equal to N.

For reasons I shall explain below, it is convenient to write $n_i = \frac{1}{2}(1 + s_i)$, where the new variables s_i take the values ± 1. Also, if interactions are appreciable only over distances much shorter than the size of the whole system, then we can write

$$\sum_j W_{ij} = \sum_j W_{ji} = W_0 \quad (10.97)$$

where W_0 is independent of the location of cell i. This will be true except for cells close to the boundaries of the system. In the thermodynamic limit, these boundary cells will be insignificant, because their number grows with the volume only as $V^{2/3}$, while the number of interior cells is proportional to V. We now use (10.25) to construct the grand canonical partition function. It is

$$Z_{\text{gr}}(\beta, V, \mu) = \mathcal{N} \sum_{\{s_i\}} \exp\left[\tfrac{1}{2}\beta\bar{\mu} \sum_i s_i - \tfrac{1}{8}\beta \sum_{i,j} W_{ij} s_i s_j\right] \quad (10.98)$$

where the modified chemical potential is given by

$$\beta\bar{\mu} = \beta\mu + \ln\left[\left(\frac{2\pi m}{\beta h^2}\right)^{3/2} v\right] - \tfrac{1}{2}\beta W_0 \quad (10.99)$$

and the factor $\mathcal{N}$, which is independent of the s_i, is

$$\mathcal{N} = \exp\left[-\left(\frac{\beta V}{2v}\right)(\bar{\mu} + \tfrac{1}{4}W_0)\right]. \quad (10.100)$$

The special value of this result is that, apart from the factor $\mathcal{N}$, it has the same form as the partition function of a well-known model for ferromagnetism, the *Ising model*, which we shall encounter in the next chapter. In that model, the variables s_i represent atomic spins (or magnetic dipole moments) situated at the sites of a crystal lattice. That this analogy between a ferromagnet and an imperfect gas can be made is, as we shall see, both theoretically important and experimentally well verified.

10.9 Analogies Between Field Theory and Statistical Mechanics

Since both quantum mechanics and statistical mechanics require us to calculate suitably weighted averages of physical quantities, it is not too surprising that formal analogies can be made between them. Under appropriate circumstances, however, these analogies can be closer than we might have expected, and it is interesting to see how they work out. Consider first of all the imaginary-time action (10.76) for a scalar field theory at finite temperature. In the integrand, the imaginary time variable appears on an equal footing with the spatial coordinates so, in effect, $\phi(\boldsymbol{x}, \tau)$ lives in a $(d+1)$-dimensional Euclidean space, d being the original number of spatial dimensions, which has a finite extent β in the extra dimension. The extra dimension is sometimes regarded as having a quantum-mechanical origin, in the following sense. The Hamiltonian of the scalar field theory may be written as

$$\hat{H} = \int \mathrm{d}^3x \left[\hat{\Pi}^\dagger \hat{\Pi} + \nabla\hat{\phi}^\dagger \cdot \nabla\hat{\phi} + m^2\hat{\phi}^\dagger\hat{\phi} + \tfrac{1}{4}\lambda(\hat{\phi}^\dagger\hat{\phi})^2\right] \tag{10.101}$$

and may loosely be compared with (10.94) for a classical gas. In the classical case, the momenta can be trivially integrated out leaving, as in (10.95), a configurational integral involving the potential energy part of the Hamiltonian with its original number of dimensions. By contrast, we could regard (10.76) as being obtained from (10.101) by again dropping the momentum term, but now adding an extra spatial dimension. If (10.101) were to be interpreted not as the Hamiltonian of a quantum field theory, but as a classical Hamiltonian (with ϕ being, say, the displacement of a continuous vibrating medium), then the configurational integral would be weighted with the exponential of $-\beta$ times its potential energy part. It would, in other words, be similar to (10.80), but with a factor β in the exponent instead of the integral over an extra dimension. While we must obviously be cautious when arguing in this way, it is frequently true that the properties of a quantum-mechanical system in d dimensions can be related to those of a $(d+1)$-dimensional classical system.

There is clearly also an analogy between the configurational integrals of classical statistical mechanics and the functional integrals of chapter 9, which represent purely quantum-mechanical expectation values. If, in a functional integral such as (9.32), we make the replacement $t = -\mathrm{i}x^4$, the weight function becomes $\exp(-S_{\mathrm{E}})$, where the Euclidean action is

$$S_{\mathrm{E}} = \int \mathrm{d}^4x[\nabla\phi^* \cdot \nabla\phi + m^2\phi^*\phi + \tfrac{1}{4}\lambda(\phi^*\phi)^2] \tag{10.102}$$

the gradient operator ∇ now being the four-dimensional Euclidean one. The introduction of an imaginary time here has nothing to do with temperature—the fourth Euclidean dimension being of infinite extent—and is, in fact, equivalent to the Wick rotation we used to evaluate Feynman integrals such as (9.76). The original Lorentz invariance of the action has been replaced by invariance under rotations in four-dimensional Euclidean space. We see that there is a rough correspondence between the Euclidean functional integral and the configurational integrals or sums of classical statistical mechanics, if we make S_{E} correspond to βW, W being the potential energy.

For a sum like (10.98), this correspondence can be made more precise by a change of variables known as the *Hubbard–Stratonovich transformation*. If we denote by Γ_{ij} the inverse of the matrix $(-\frac{1}{4}W_{ij})$, we can prove the identity

$$\exp\left[\tfrac{1}{2}\beta\sum_{i,j}(-\tfrac{1}{4}W_{ij})s_i s_j\right] = Q\int_{-\infty}^{\infty}\prod_i \mathrm{d}\Phi_i \exp\left[-\frac{1}{2\beta}\sum_{i,j}\Gamma_{ij}\Phi_i\Phi_j + \sum_i \Phi_i s_i\right] \tag{10.103}$$

by completing the square on the right-hand side; that is, by making the shift $\Phi_i \to \Phi_i - \frac{1}{4}\beta\sum_j W_{ij}s_j$. Obviously, Q is the appropriate normalizing factor. Applying this identity to the partition function (10.98), it becomes easy to carry out the sums over the s_i:

$$\sum_{s=\pm 1}\exp\left[(\tfrac{1}{2}\beta\bar{\mu} + \Phi)s\right] = 2\cosh(\tfrac{1}{2}\beta\bar{\mu} + \Phi). \tag{10.104}$$

Thus, the partition function of the lattice gas becomes

$$Z_{\mathrm{gr}} = \bar{Z}\int_{-\infty}^{\infty}\prod_i \mathrm{d}\Phi_i \exp\left[-\frac{1}{2\beta}\sum_{i,j}\Gamma_{ij}\Phi_i\Phi_j + \sum_i \ln\cosh(\tfrac{1}{2}\beta\bar{\mu} + \Phi_i)\right] \tag{10.105}$$

where $\bar{Z}$ denotes the collection of normalizing factors we have accumulated. This partition function will be essentially identical to a functional integral if we take the cells of the lattice, positioned, say, at the points $\boldsymbol{x}_i$, to be tiny compared with the total size of the system. Let us, indeed, regard the variables Φ_i as the values of a continuous function $\Phi(\boldsymbol{x})$ at the points $\boldsymbol{x} = \boldsymbol{x}_i$. Correspondingly, we would like to convert the sums over lattice sites into integrals, using the replacement $\sum_i \to v^{-1}\int \mathrm{d}^3x$, but we have the matrix Γ_{ij} to contend with. This matrix can be regarded as a function of the distance between two lattice sites, say $\Gamma_{ij} = \Gamma(|\boldsymbol{x}_i - \boldsymbol{x}_j|)$, and under some circumstances it is permissible to expand its Fourier transform as

$$\Gamma_{ij} = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)}\left[\Gamma_0 + \Gamma_1 k^2 + \ldots\right] \tag{10.106}$$

keeping only the first two terms. On account of the Fourier representation of the Dirac δ function, we can rewrite this as

$$\Gamma_{ij} = \left[\Gamma_0 + \Gamma_1 \nabla_i \cdot \nabla_j + \ldots\right] \delta(\boldsymbol{x}_i - \boldsymbol{x}_j). \qquad (10.107)$$

In this way, we can approximate (10.105) by the functional integral

$$Z_{\text{gr}} \approx \int \mathcal{D}\Phi(\boldsymbol{x}) \, \exp\left[-\mathcal{H}(\Phi)\right] \qquad (10.108)$$

where the effective 'reduced Hamiltonian' is

$$\mathcal{H}(\Phi) = \int \mathrm{d}^3x \left[\left(\frac{\Gamma_1}{2\beta v^2}\right)\nabla\Phi\cdot\nabla\Phi + \left(\frac{\Gamma_0}{2\beta v^2}\right)\Phi^2 - \frac{1}{v}\ln\cosh\left(\tfrac{1}{2}\beta\bar{\mu} + \Phi\right)\right]. \qquad (10.109)$$

To derive this form, I have used integrations by parts to make the derivatives in (10.107) act on Φ and used the δ function to do one of the space integrals. A final change of the integration variable

$$\Phi(\boldsymbol{x}) = -\tfrac{1}{2}\beta\bar{\mu} + (\beta v^2/\Gamma_1)^{1/2}\phi(\boldsymbol{x}) \qquad (10.110)$$

together with the expansion $\ln\cosh(y) = \frac{1}{2}y^2 - \frac{1}{12}y^4 + \ldots$ enables us to write

$$\mathcal{H} \approx \int \mathrm{d}^3x \left[\tfrac{1}{2}\nabla\phi\cdot\nabla\phi + \tfrac{1}{2}m^2\phi^2 + \tfrac{1}{4}\lambda\phi^4 - J\phi\right] \qquad (10.111)$$

where

$$m^2 = (\Gamma_0 - \beta v)/\Gamma_1 \qquad \lambda = \beta^3 v^3/3\Gamma_1^2 \qquad J = (\beta\Gamma_0^2/4\Gamma_1 v^2)^{1/2}\bar{\mu} \quad (10.112)$$

which is equivalent to the Euclidean action for a relativistic scalar field. (The factors of $\frac{1}{2}$ in the first two terms give the right normalization for a real field, as shown by exercise 7.1.) The equivalence we have derived is, of course, only approximate. It will be valid, roughly speaking, when the functions $\phi(\boldsymbol{x})$ that make the most important contributions to the functional integral are small (so that higher-order terms in the expansion of $\ln\cosh(y)$ can be neglected) and vary slowly with spatial position (so that the higher-order terms in the gradient expansion (10.107) can be neglected). As will transpire in the next chapter, these approximations are well justified in the neighbourhood of a *critical point*, where the analogy is most useful. It will be noticed that (10.102) and (10.111) involve different numbers of spatial dimensions and we shall see that this has important consequences.

Exercises

10.1. Consider a classical one-dimensional harmonic oscillator, with Hamiltonian $H = p^2/2m + m\omega^2 x^2/2$. What are the curves of constant energy in its two-dimensional phase space? Show that $\Sigma(E) = 2\pi/\omega$. Show that both the long-time average and the microcanonical average of a function $f(x, p)$ are given by

$$\frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta\, f\left((2E/m\omega^2)^{1/2}\sin\theta,\, (2mE)^{1/2}\cos\theta\right).$$

This system is therefore ergodic. By considering the flow of an ensemble of points on the energy surface, show that it is *not* mixing.

10.2. Consider a single classical, non-relativistic particle of mass m in a cubical box of side L, which rebounds elastically each time it collides with a wall. Suppose that two opposite walls lie in the planes $x = 0$ and $x = L$. Show that, averaged over a long period of time, the momentum per unit time that the particle exchanges with each of these walls is mv_x^2/L. Hence show that the pressure exerted by a gas of N particles is $p = Nm\langle v^2\rangle/3L^3$, where v is the magnitude of the velocity of any one particle and the average is over all the particles, and verify the relation $p = 2U/3V$. Note that this result does not assume any particular distribution of velocities.

10.3. For an open system, define the fluctuation ΔN in the number of particles by $(\Delta N)^2 = \langle(N - \bar{N})^2\rangle$. Show that $(\Delta N)^2 = \partial^2 \ln Z_{\mathrm{gr}}/\partial(\beta\mu)^2$. For a classical ideal gas, show that $\Delta N/\bar{N} = \bar{N}^{-1/2}$. In the same way, show that the relative fluctuations in the internal energy U are proportional to $\bar{N}^{-1/2}$.

10.4. The partition function for the *pressure ensemble* (or *isobaric* ensemble) is

$$Z_{\mathrm{pr}}(\beta, p, N) = \int_0^\infty \mathrm{d}V\, \mathrm{e}^{-\beta pV} Z_{\mathrm{can}}(\beta, V, N).$$

Calculate this partition function for a classical ideal gas. Suggest an expression, in terms of Z_{pr} and its derivatives, for the mean volume of a system maintained at constant pressure p, and check it by recovering the ideal gas equation of state in the thermodynamic limit. Show that, in the thermodynamic limit, the quantity $G = -k_{\mathrm{B}}T \ln Z_{\mathrm{pr}}$ is the *Gibbs free energy* $G = F + pV$. Show that the chemical potential μ is the Gibbs free energy per particle.

10.5. Given any set of objects, mathematicians define an *equivalence relation* $\sim$ between any two of them as a relation that has the three properties:

(i) for each object a in the set, $a \sim a$ (reflexivity);
(ii) if $a \sim b$ then $b \sim a$ (symmetry);
(iii) if $a \sim b$ and $b \sim c$, then $a \sim c$ (transitivity).

Show that these properties allow one to divide the set into *equivalence classes* such that all members of any one class are ‘equivalent’ to each other and no two objects belonging to different classes are ‘equivalent’ to each other.

Consider a set of macroscopic physical systems, and interpret $a \sim b$ to mean ‘a has the same temperature as b’. How is the zeroth law of thermodynamics relevant to the possibility of assigning unique temperatures to these systems?

10.6. Show that the density operator (10.50) is Hermitian and that the trace in (10.52) does not depend on which complete orthonormal set of states is used to compute it.

10.7. In the complex z plane, let C be the closed contour which runs from $-\infty$ to $+\infty$ just below the real axis and returns to $-\infty$ just above the real axis. Show that, for any sufficiently well-behaved function $f(z)$,

$$\lim_{\epsilon\to 0}\oint_{\mathrm{C}} \mathrm{d}z \frac{\mathrm{e}^{\mathrm{i}\epsilon z} f(z)}{\mathrm{e}^{2\pi \mathrm{i}z}-1} = \sum_{n=-\infty}^{\infty} f(n).$$

Verify (10.82) by choosing $f(z) = (z^2 + a^2)^{-1}$ and deforming the contour in an appropriate manner.

10.8. Consider the field-theoretic partition function (10.84) in the limit that βm is very large and ignoring the last (vacuum energy) term. By making the change of variable $x \to (2\beta m)^{1/2}x$, show that (10.84) reduces to the non-relativistic partition function (10.65) for spin-0 particles, with $\hbar = 1$ and a chemical potential $\mu = -m$.

10.9. Consider a gas of N hydrogen atoms in a container of volume V, at a temperature high enough for all H_2 molecules to be dissociated and some atoms to be ionized. Using classical, non-relativistic statistical mechanics, work out the canonical partition function for $N-\nu$ indistinguishable atoms, ν indistinguishable protons and ν indistinguishable electrons. For each ionized atom, include a potential energy I, equal to the ionization potential. Assume that the masses of a hydrogen atom and a proton are equal. By finding the most probable value of ν, show that the fraction $x = \nu/N$ of ionized atoms is given by the *Saha equation*

$$\frac{x^2}{1-x} = \frac{1}{n}\left(\frac{2\pi m}{\beta h^2}\right)^{3/2} \mathrm{e}^{-\beta I}$$

where m is the electron mass and $n = N/V$. Note that this result depends on h, which is an arbitrary parameter in the classical theory. Why is this? Why would you expect to obtain the correct answer by taking h to be Planck's constant?

10.10. From equations (10.88) and (10.91)–(10.93), it might appear that a fermion simply counts as $\frac{7}{8}$ of a boson as far as black-body radiation is concerned, but this is not so. By direct calculation or informal arguments, convince yourself that the number density of species i is given by

$$\frac{N}{V} = \frac{g_i}{2\pi^2\beta^3}\int_0^{\infty} \mathrm{d}x\, x^2(\mathrm{e}^x \pm 1)^{-1}.$$

Show that the fermionic integral is $\frac{3}{4}$ of the bosonic one. The value of the bosonic integral is $2\zeta(3)$, where ζ is the Riemann zeta function.

Chapter 11

Phase Transitions

Among the many applications of statistical mechanics, some of the most intriguing and challenging theoretical problems arise in connection with *phase transitions*. These are abrupt changes of state such as occur, for example, when a liquid is transformed into a vapour, a ferromagnet loses its magnetization upon heating to its Curie temperature, or at the onset at sufficiently low temperatures of superfluidity or superconductivity. It is within the theory of phase transitions, too, that the mathematical relationships between statistical mechanics and relativistic field theories are most powerful. Indeed, the idea of *spontaneous symmetry breaking*, which lies at the heart of the theory of phase transitions, is the crucial ingredient that turns the gauge theories of chapter 8 into a real working model of the fundamental forces of nature, to be discussed in the next chapter.

It is not possible in the space of a single chapter to cover adequately the wide and diverse range of phenomena that theoretical and experimental ingenuity have uncovered. I shall therefore discuss only a few standard examples and the key theoretical arguments that have been devised to deal with them. In almost all cases, phase transitions can occur only by virtue of interactions between particles. This, indeed, is what gives rise to the greatest theoretical challenges. The one exception to this rule is the case of *Bose–Einstein condensation* in an ideal Bose gas, which I shall discuss first. The greater part of the chapter will deal with the gas–liquid and ferromagnetic transitions, which illustrate most of the essential theory, and I shall end by describing the Ginzburg–Landau theory of superconductivity, which provides the closest analogy with the gauge theories of particle physics.

11.1 Bose–Einstein Condensation

Consider an ideal gas of spin-0 particles. According to (10.64), the average number of particles in the ith momentum state, with momentum given by (10.61) is

$$n_i = z[\exp(\beta\epsilon_i) - z]^{-1}. \tag{11.1}$$

For a given number of particles per unit volume, the fugacity z is determined implicitly in terms of $\bar{N}/V$ and temperature by an equation of the form (10.66). By its definition (10.24), z is positive. On the other hand, since the occupation numbers (11.1) cannot be negative, z cannot be greater than $\exp(\beta\epsilon_0)$, where ϵ_0 is the smallest single-particle energy. For a large volume, we can take this energy to be zero, so $0 < z < 1$, which means that the chemical potential μ must be negative. The interesting question is, what happens as z approaches 1? We see from (11.1) that the occupation number of the zero-energy state can become indefinitely large. In fact, the growth of this number is limited by the total number of particles available, but it can be a significant fraction of the total number. This phenomenon, known as *Bose–Einstein condensation*, is the basic cause of superfluidity and superconductivity.

When the zero-energy state is macroscopically occupied, we have to reconsider equations such as (10.65) and (10.66), where we replaced a sum over momentum states with an integral. This is normally valid because the momentum eigenvalues are very closely spaced, but it assumes that the fraction of particles with momentum in the infinitesimal range d^3p is infinitesimal. When there is condensation, this will not be true for the element d^3p which includes the zero-energy state. In fact, the integrals in (10.65) and (10.66) do assign only an infinitesimal fraction of the particles to this element, so we can correct them simply by adding on the contributions of the condensed particles. For the grand potential and particle number per unit volume, we obtain

$$\frac{\Omega}{V} = \frac{1}{\beta V}\ln(1-z) + 4\pi\beta^{-5/2}\left(\frac{2m}{h^2}\right)^{3/2}\int_0^\infty \mathrm{d}x\, x^2 \ln(1 - z\mathrm{e}^{-x^2}) \tag{11.2}$$

$$\frac{\bar{N}}{V} = \frac{\bar{n}_0}{V} + 4\pi\left(\frac{2m}{\beta h^2}\right)^{3/2} z\int_0^\infty \mathrm{d}x\, x^2\mathrm{e}^{-x^2}(1 - z\mathrm{e}^{-x^2})^{-1} \tag{11.3}$$

where $\bar{n}_0$ is the average number of condensed particles. These equations are to be understood as applying to the thermodynamic limit. When $V \to \infty$, the condensation terms go to zero, unless z is infinitesimally close to 1 and the number of condensed particles per unit volume is finite.

The conditions under which the condensation occurs can be investigated as follows. Suppose first that condensation does occur. Then z is infinitesimally close to 1. The second term in (11.3) is proportional to the integral

$$4\pi^{-1/2}\int_0^\infty \mathrm{d}x\, x^2(\mathrm{e}^{x^2} - 1)^{-1} = \zeta(\tfrac{3}{2}) = 2.612\ldots \tag{11.4}$$

where ζ is the Riemann zeta function, and we have

$$\frac{\bar{N}}{V} = \frac{\bar{n}_0}{V} + 2.612\left(\frac{2\pi m k_\mathrm{B}}{h^2}\right)^{3/2} T^{3/2}. \tag{11.5}$$

For a given number density, there is a *critical temperature* T_c at which the number of condensed particles just vanishes:

$$T_c = \frac{h^2}{2\pi m k_B}\left(\frac{\bar{N}}{2.612\,V}\right)^{2/3}. \tag{11.6}$$

At temperatures lower than this, $\bar{n}_0$ is a non-zero fraction of $\bar{N}$. At higher temperatures, on the other hand, (11.5) cannot be true, because $\bar{n}_0$ cannot be negative. For $T > T_c$, therefore, the assumption that $\bar{n}_0$ is macroscopically large is not self-consistent; we must have $\bar{n}_0/V \to 0$ in the thermodynamic limit and z must be smaller than 1.

In the condensed phase (that is, the low-temperature state in which condensation occurs), it is easy to see that the fraction of particles in the condensate is

$$\frac{\bar{n}_0}{\bar{N}} = 1 - \left(\frac{T}{T_c}\right)^{3/2}. \tag{11.7}$$

Under the influence of an applied force, such as gravity or the attraction of container walls, this condensate moves as a coherent whole and is responsible for the frictionless flow characteristic of superfluid helium. (Helium is the only substance known to exhibit superfluidity. It is not, however, an ideal gas and intermolecular forces are essential for understanding its properties in detail.) The condensate can be described by a macroscopic wavefunction ϕ, whose magnitude is proportional to $\sqrt{\bar{n}_0}$. The temperature dependence of quantities like $|\phi|$ in the immediate neighbourhood of a critical temperature will be a recurring theme. If we expand (11.7) in powers of $T - T_c$, we find

$$|\phi| \sim (T_c - T)^{\beta} \tag{11.8}$$

where β, an example of what is called a *critical exponent*, has the value $\frac{1}{2}$. The symbol $\sim$ indicates both that a constant of proportionality is missing and that this is only the leading behaviour when $T_c - T$ is small.

Another important feature, which is common to all phase transitions, is that the transition is sharply defined only in the thermodynamic limit. When $V \to \infty$ in (11.3), we can draw a sharp distinction between the condensed phase, in which $\bar{n}_0/V$ has a non-zero limit, and the normal phase in which it goes to zero. When the volume is large but finite, there is a narrow range of temperature in which $\bar{n}_0/V$ decreases from being a significant fraction of $\bar{N}/V$ to being extremely small, but no precise dividing line between the two phases. Although experiments deal with finite systems, these systems occupy a volume that is extremely large compared with average intermolecular distances. Under these circumstances, the theoretical ambiguity as to the precise location of a critical temperature may well be much smaller than the resolution in temperature that an experimenter can achieve. Thus, to all intents and purposes, well defined phase transitions can indeed be observed in practice.

11.2 Critical Points in Fluids and Magnets

Much of the theoretical interest in phase transitions has to do with *critical points*. The exact nature of a critical point will emerge as we study examples, but one essential feature is already apparent from the case of Bose–Einstein condensation. The condensed phase, which exists below T_c, is distinguished from the normal, high-temperature phase by a non-zero value of $\bar{n}_0/V$. On approaching the critical temperature, this quantity goes continuously to zero, and so, exactly *at* the critical temperature, the condensed and normal phases are identical. This behaviour is distinctive of critical points, which may also be described as *continuous* or *second-order* phase transitions, the terminology depending somewhat upon its context. Had $\bar{n}_0/V$ dropped discontinuously to zero at T_c, it would have been possible for distinct condensed and normal phases to coexist with each other at T_c, which is characteristic of a *first-order* phase transition. A classification of phase transitions due to P Ehrenfest defines a phase transition to be of *n*th order if an *n*th derivative of the appropriate thermodynamic potential is discontinuous, while all of its $(n-1)$th derivatives are continuous. If we introduce a separate chemical potential μ_0 for particles in the zero-energy state, then $\bar{n}_0$ is the first derivative of Ω with respect to μ_0. It is continuous at T_c, but $\partial n_0/\partial T$ is not, so the condensation is indeed second-order according to this classification. However, the singularities found at phase transitions are often more complicated than simple discontinuities, so the general classification scheme has fallen out of common use.

Two standard, easily studied examples of critical points are those which occur in simple fluids and in ferromagnets, and I shall deal first with ferromagnetism. As readers are no doubt aware, a permanently magnetized sample of, say, iron typically contains a number of domains, the directions of magnetization being different in neighbouring domains. The physical factors that control the size of a domain have no direct bearing on the phase transitions we are discussing, so I shall simplify matters by assuming that the magnetization of the sample is completely uniform. In practice, our considerations will apply to the interior of a single domain. The magnetization M_S that exists in the absence of any applied magnetic field is called the *spontaneous magnetization* and its magnitude depends on temperature in the manner sketched in figure 11.1. Upon heating to the critical (or Curie) temperature T_c, the spontaneous magnetization vanishes continuously. In the immediate neighbourhood of T_c, called the *critical region*, we find

$$M_S \sim (T_c - T)^{\beta}. \tag{11.9}$$

The exponent β varies rather little from one ferromagnetic material to another and is typically about $\frac{1}{3}$.

The direction in which the magnetization points usually lies along one of several *easy axes*, defined by the crystal structure of the material. For simplicity, I shall consider only *uniaxial* materials in which there is only one easy axis. Then the magnetization can point in one of two opposite directions along this axis. Consider what happens when a magnetic field H is applied in a direction parallel

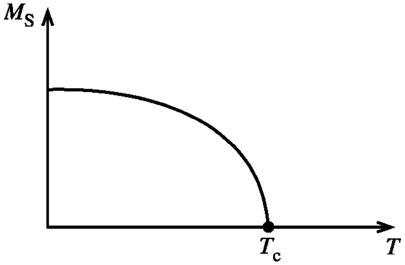


Figure 11.1. The spontaneous magnetization of a ferromagnet as a function of temperature.

to the easy axis. To be specific, let us suppose that the magnet is heated to a temperature above T_c, at which point a large magnetic field is applied. There will be a magnetization parallel to H, which would decrease to zero if the field were removed. If, in the presence of H, the magnet is cooled to below T_c, the magnetization remains parallel to H, but reduces to M_S when the field is removed. If the process is repeated, but with the direction of H reversed, we end up with a magnetization of magnitude M_S pointing in the opposite direction.

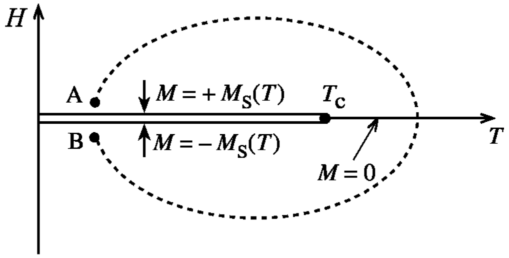


Figure 11.2. Phase diagram of a ferromagnet in the temperature-magnetic field plane. Below T_c, the magnetization is discontinuous at $H = 0$. However, by varying H and T along the broken curve, we can pass from state A to state B without encountering a phase transition.

Ideally, given a temperature T and magnetic field H, there is a magnetization $M(T, H)$ which has a unique value, except when $H = 0$ for $T < T_c$, where the limit as $H \to 0$ of $M(T, H)$ is either $M_S(T)$ if H is positive or $-M_S(T)$ if H is negative. We may therefore draw an idealized *phase diagram* as in figure 11.2. As far as the line $H = 0$ is concerned, we can identify three different phases, namely two ferromagnetic phases, distinguished by oppositely directed magnetizations, which exist below T_c, and the *paramagnetic* phase, with $M = 0$, above T_c. At the critical point $(T, H) = (T_c, 0)$, the two ferromagnetic phases become identical and also indistinguishable from the paramagnetic phase. The line $H = 0$, $T < T_c$ is a line of two-phase coexistence, where oppositely magnetized domains can coexist in the same sample. Ideally, it is a line of first-

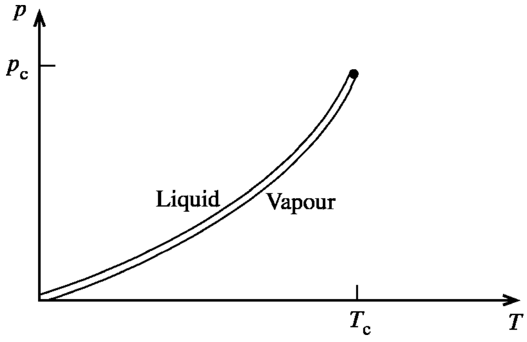


Figure 11.3. Phase diagram of a simple fluid in the temperature–pressure plane.

order phase transitions, since the magnetization changes discontinuously from M_S to $-M_S$ as the magnetic field decreases through zero. We see, however, that any two states, say A and B in figure 11.2, can be connected by a path along which no phase transition occurs. The essential definition of a critical point is that it marks the end of a line where two or more phases coexist, and that these phases become identical in a continuous manner. In practice, the way in which the magnetization of a sample varies with temperature and magnetic field is more complicated and it is necessary to consider, for example, the motion of domain walls, which gives rise to hysteresis. The actual magnetization of a sample is not given by a single-valued function, but depends on its history. Nevertheless, the function $M(T, H)$ can be found by careful experimental procedures and it is this function that we hope to be able to calculate, at least approximately, from equilibrium statistical mechanics.

The *magnetic susceptibility* is defined by

$$\chi = \partial M/\partial H. \tag{11.10}$$

The zero-field susceptibility, sometimes called the *initial susceptibility*, is found to diverge at the critical point. That is, it becomes infinite, and it does so as a power of $|T - T_c|$:

$$\chi(T, 0) \approx \chi_0 |T - T_c|^{-\gamma}. \tag{11.11}$$

The critical exponent γ, which has similar values of about 1.3 for all ferromagnets, is found to be the same, whether the critical temperature is approached from above or below, but the amplitude χ_0 may be different in the two cases.

The behaviour of simple fluids is quite analogous to that of ferromagnets. Figure 11.3 represents the vapour–pressure curve $p = p_v(T)$, which ends at a critical point (T_c, p_c). By speaking of a 'simple fluid', I mean that additional complications, such as the possibility of solidification, will be ignored. Although most real substances have more complicated phase diagrams than the one shown

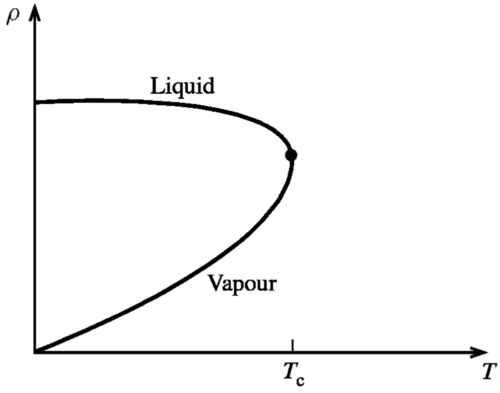


Figure 11.4. Variation with temperature of the densities of liquid and vapour phases of a simple fluid at the vapour pressure.

in figure 11.3, these complications do not affect the critical properties we are discussing. Along the vapour–pressure curve, the liquid and vapour phases of the same substance can coexist in the same container. By varying the pressure at a fixed temperature below T_c, we can transform liquid into vapour or *vice versa*. This is a first-order transition, because the density ρ changes discontinuously. If we plot the densities of the liquid and vapour, both measured at the vapour pressure, as functions of temperature, the result is that sketched in figure 11.4. It is obviously analogous to the spontaneous magnetization curve of figure 11.1, if we include the oppositely directed magnetization, except that it is not symmetrical. Near the critical point, the difference in density between the liquid and vapour is found to vary as

$$\rho_\ell - \rho_v \sim (T_c - T)^\beta. \tag{11.12}$$

Measured values of the exponent β are very similar for all fluids. Remarkably, they are also very similar to the values obtained for ferromagnets, being in the neighbourhood of $\frac{1}{3}$. Indeed, it is found that all *critical phenomena* (that is, the properties of systems in the neighbourhoods of their critical points) are substantially independent of the detailed microscopic constitution of the system considered. This *universality* of critical phenomena is, of course, one of the principal features that we should like to understand theoretically.

It is convenient to focus theoretical discussions on magnetic systems because, as is evident from figures 11.2 and 11.3, they possess a greater degree of symmetry. The magnetization of a macroscopic sample is a magnetic dipole moment per unit volume, which may have contributions from the intrinsic dipole moments of fixed atoms or ions and mobile electrons and also from the orbital motion of electrons. When the major contribution is from mobile electrons, the magnetism is said to be *itinerant*. When the major contribution is from atoms or ions fixed at the sites of a crystal lattice or from electrons which, though mobile, tend to congregate near these lattice sites, the magnetization is said to be *localized.* The exact degree of itineracy or localization is not easy to establish, but it appears that the three common metallic ferromagnets, namely iron, cobalt and nickel, are predominantly itinerant. Theoretically, it is somewhat easier to deal with localized magnets and, because of universality, this does not, in the end, make much difference as far as their critical properties are concerned. I shall therefore regard ferromagnetism as arising from localized moments situated at the sites of a lattice. Each of these magnetic moments is proportional to the intrinsic spin of an atom or ion, and the basic constituents of the magnet are conventionally referred to as *spins*.

To understand the origin of universality, it is necessary to consider correlations between the directions of spins at different sites. Our sample will exhibit a net magnetization if, on average, all the spins tend to point in the same direction. In a large sample, the average of a spin variable $\boldsymbol{s}_i$ at the ith lattice site will be independent of the particular site. The magnetization per spin is

$$\boldsymbol{M} = m\langle \boldsymbol{s}_i \rangle \tag{11.13}$$

where $m\boldsymbol{s}_i$ is the magnetic moment associated with the spin. The fluctuations of a given spin away from its average value are measured by $\boldsymbol{s}_i - \langle \boldsymbol{s}_i \rangle$. What particularly concerns us is the correlation between such fluctuations at two different sites. We define the *correlation function* $G(\boldsymbol{r}_i - \boldsymbol{r}_j)$ as

$$G(\boldsymbol{r}_i - \boldsymbol{r}_j) = \langle (\boldsymbol{s}_i - \langle \boldsymbol{s}_i \rangle) \cdot (\boldsymbol{s}_j - \langle \boldsymbol{s}_j \rangle) \rangle = \langle \boldsymbol{s}_i \cdot \boldsymbol{s}_j \rangle - \langle \boldsymbol{s}_i \rangle \cdot \langle \boldsymbol{s}_j \rangle \tag{11.14}$$

where $\boldsymbol{r}_i$ is the position of the ith lattice site. Analogous correlation functions can be defined in terms of magnetization density for itinerant magnets or density fluctuations in a fluid. Assuming that only short-ranged forces act between spins, we would expect this correlation function to decay to zero at large distances, and so it does. Under most circumstances, we find

$$G(\boldsymbol{r}_i - \boldsymbol{r}_j) \sim \exp(-|\boldsymbol{r}_i - \boldsymbol{r}_j|/\xi) \tag{11.15}$$

where ξ is a characteristic distance called the *correlation length*. The correlation length depends on temperature and on the applied magnetic field. In the absence of an applied field, it diverges at the critical point, and this divergence is governed by a new critical exponent ν:

$$\xi(T) \approx \xi_0 |T - T_\mathrm{c}|^{-\nu}. \tag{11.16}$$

As with the susceptibility, the same exponent governs the divergence as the critical temperature is approached from above or from below, but the amplitudes ξ_0 may be different. Typically, we find $\nu \approx 0.6$–0.7.

This divergence of the correlation length is at the root of the universality of critical phenomena. Because fluctuations are strongly correlated over large distances, they, and the critical properties that depend on them, are insensitive to details of the forces that act over microscopic distances. Experimentally, the correlation functions can be investigated by scattering. In the case of magnets, the scattering of neutrons is affected by magnetic forces, while the scattering of light by fluids depends on density correlations. When the correlation length is large, the scattered waves from points widely separated in the sample are coherent, and so strong scattering results. This is visible to the naked eye in a fluid near its critical point. The strong scattering by a substance which is normally transparent gives it a foggy or milky appearance known as *critical opalescence*. From a theoretical point of view, we might expect that quantities such as critical exponents could be calculated on the basis of quite highly idealized models, which take little account of the detailed microscopic constitution of real materials, and this appears to be borne out in practice. Later on, we shall see in rather more detail why this is so.

11.3 The Ising Model and its Approximation by a Field Theory

The forces that tend to align spins in a ferromagnet have electrostatic and quantum-mechanical origins. For example, if the spins of two electrons (which are fermions) in neighbouring atoms are in a triplet state, which roughly means that they are parallel, then, to maintain the overall antisymmetry of the two-electron state, their orbital motion must be described by an antisymmetric combination of atomic orbitals. Conversely, if their spins are in a singlet, antiparallel state, then their orbital state must be symmetric. The expectation value of the electrons' electrostatic energy is different in the symmetric and antisymmetric orbital states, and therefore also in the singlet and triplet spin states. This leads to an effective interaction between spins, called an *exchange* interaction. To study magnetic effects, we would like to use an effective Hamiltonian that depends only on spin degrees of freedom. It was shown by Heisenberg that such a Hamiltonian must have the form

$$\hat{H} = -\sum_{i,j} J_{ij} \hat{\boldsymbol{S}}_i \cdot \hat{\boldsymbol{S}}_j \tag{11.17}$$

where $\hat{\boldsymbol{S}}_i$ is the spin operator for the ith lattice site and J_{ij} is a symmetric matrix of constants representing the exchange energies. Usually, these energies will be appreciable only when sites i and j are close together. The exchange energies can have either sign. If the J_{ij} are predominantly positive, then parallel spins have the lower energy and ferromagnetism will result; if they are negative, then we shall have antiferromagnetism.

In a uniaxial ferromagnet, the spins point preferentially along one crystallographic axis, say the z axis, so we can delete the x and y spin components in (11.17). In that case, all the remaining operators commute with each other, and we can choose a set of basis states in which they are all diagonal. If we take the $\hat{S}_i$ to be spin-$\frac{1}{2}$ operators, their eigenvalues are $\pm\frac{1}{2}\hbar$. For theoretical purposes, it is useful to imagine that an independent magnetic field H_i can be applied at each lattice site. It is also convenient to absorb factors of $\frac{1}{2}\hbar$ into the definitions of J_{ij} and H_i, and the magnetic moment m in (11.13) into the definition of H_i. Then the partition function may be written as

$$Z(\beta, \{H_i\}) = \mathrm{Tr}\ \mathrm{e}^{-\beta\hat{H}} = \sum_{\{s_i=\pm 1\}} \mathrm{e}^{-\beta H_\mathrm{I}} \tag{11.18}$$

where

$$H_\mathrm{I} = -\sum_{i,j} J_{ij} s_i s_j - \sum_i H_i s_i. \tag{11.19}$$

This is the *Ising model.* As a model of a ferromagnet, it is clearly rather idealized, taking into account only the configurational average of spin degrees of freedom. Thus, the free energy $F = -k_\mathrm{B}T \ln Z$ obtained from (11.18) represents not the whole free energy of a ferromagnetic material, but only that contribution to it which is directly involved in the ferromagnetic transition.

The partition function (11.18) is obviously of the same form as the configurational sum in (10.98) for the lattice gas, so long as we identify the adjusted chemical potential (10.99) with a uniform magnetic field. We see from (10.98) that the grand potential of the lattice gas receives a contribution from the factor $\mathcal{N}$ multiplying the configuration sum, but this contribution is a smooth function of temperature and chemical potential and cannot be directly involved in the gas–liquid transition.

From (11.18), the correlation function (11.14) can be expressed as

$$G(\boldsymbol{r}_i - \boldsymbol{r}_j) = \beta^{-2} \frac{\partial^2 \ln Z}{\partial H_i \partial H_j} = \beta^{-1} \frac{\partial}{\partial H_i} \langle s_j \rangle. \tag{11.20}$$

For a uniform magnetic field, the magnetic susceptibility can be written in terms of the correlation function as

$$\chi = \frac{\partial}{\partial H} \langle s_i \rangle = \beta \sum_j G(\boldsymbol{r}_i - \boldsymbol{r}_j). \tag{11.21}$$

We saw in the last chapter that, by means of the Hubbard–Stratonovich transformation, the spin variables s_i can be replaced with a new set of variables ϕ_i, in terms of which the Ising model takes on an appearance similar to a relativistic field theory. If we replace $\frac{1}{2}\bar{\mu}$ in the partition function (10.105) with the site-dependent magnetic field H_i, we find that the averages of s_i and ϕ_i are related by

$$\langle s_i \rangle = \frac{\partial \ln Z_\mathrm{gr}}{\partial(\beta H_i)} = \langle \tanh(\beta H_i + \Phi_i) \rangle = \langle \tanh(a\phi_i) \rangle \tag{11.22}$$

where a is the factor $(\beta v^2/\Gamma_1)^{1/2}$ that appears in (10.110). Near the critical point, the magnetization is small, so it becomes legitimate to use the approximation $\tanh(a\phi) \approx a\phi$, and we can take $\langle \phi \rangle$ to be proportional to the magnetization.

Moreover, since critical phenomena are associated with strong correlations over large distances, the gradient expansion in (10.111), should be a reasonable approximation. To be explicit about this, the functions $\phi(\boldsymbol{x})$ that are most important in the functional integral (10.108) ought, near the critical point, to be those that vary significantly only over distances of the order of ξ or greater (say, $\phi(x) \sim \sin(x/\xi)$, for example). Each extra ∇ in (10.107) gives rise to an extra ∇ acting on $\phi(\boldsymbol{x})$ in the effective Hamiltonian (10.111) and hence, in effect, to a factor of ξ^{-1}, which is small. Also, near the critical point, it will be adequate to ignore the temperature dependence of the factor a in (11.22) and the parameters J and λ in (10.112), by setting $\beta = 1/k_{\mathrm{B}}T_{\mathrm{c}}$. On the other hand, the parameter m^2 vanishes when $T = v/k_{\mathrm{B}}\Gamma_0 \equiv T_0$, and we shall see that this is an approximation to the critical temperature. Near the critical temperature, we can take $m^2 \propto (T - T_0)$. These rough-and-ready arguments are made rather more precise by the renormalization-group ideas to be discussed in §11.6. In this way, we arrive at an approximate partition function for our ferromagnet, which I will now rewrite, using a notation that is traditional in this subject, as

$$Z(T, H) = \int \mathcal{D}\phi \, \exp[-H_{\mathrm{eff}}(\phi)] \tag{11.23}$$

where the effective Hamiltonian is

$$H_{\mathrm{eff}} = \int \mathrm{d}^d x \, \left(\tfrac{1}{2}\nabla\phi \cdot \nabla\phi + \tfrac{1}{2}r_0\phi^2 + \tfrac{1}{4!}u_0\phi^4 - h\phi \right). \tag{11.24}$$

The parameter r_0 is proportional to $(T - T_0)$, while h is proportional to the magnetic field H and u_0 is a constant. I have written this down as it would apply to a system that has d spatial dimensions. In practice, we normally want $d = 3$ (or sometimes $d = 1$ or $d = 2$), but we shall see that there are advantages in considering other values of d as well.

A large number of approximations stand between the field theory (11.23) and any realistic model of a ferromagnet. Nevertheless, it is believed to embody exact information about universal quantities such as the exponents β, γ and ν. However, the information we would need to calculate non-universal quantities such as the critical temperature itself or the amplitude χ_0 in (11.11) has largely been lost.

11.4 Order, Disorder and Spontaneous Symmetry Breaking

The phase transition that takes place at the Curie temperature in zero magnetic field can be described as an *order–disorder transition*. The high-temperature paramagnetic phase is one in which fluctuations in the orientation of each spin variable are entirely random, so that the configurational average is zero: the state is a disordered one. In the ferromagnetic phases, on the other hand, the spins point preferentially in one particular direction and, in this sense, the state is ordered. On

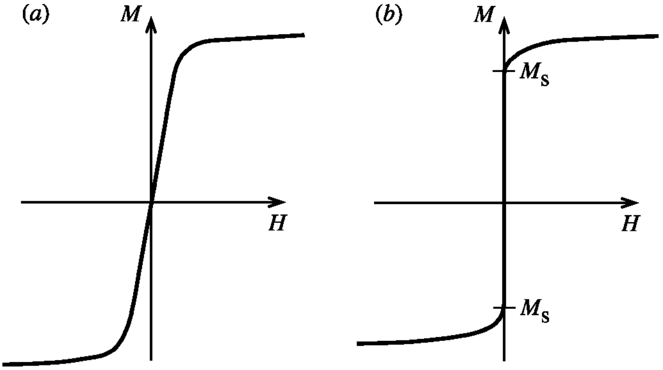


Figure 11.5. Variation of magnetization with magnetic field in a ferromagnet at a fixed temperature below $T_{\rm c}$: (*a*) finite system; (*b*) infinite system.

the face of it, it is hard to understand how such an ordered state can come about. The Ising Hamiltonian (11.19) with $H_i = 0$ has a symmetry: if we reverse the sign of every spin, it remains unchanged. Therefore, for each configuration in which a given spin s_i has the value $+1$, there is another one, obtained by reversing all the spins, in which it has the value -1, and these two configurations have the same statistical weight. Thus, the magnetization per spin ought to be zero, and this argument is apparently valid for any temperature. Indeed, for any *finite* system, the conclusion is inescapable. Within the framework of the Ising model, the only way to obtain a non-zero spontaneous magnetization is to consider an infinite system; that is, to take the thermodynamic limit.

The way in which this comes about is illustrated in figure 11.5. If we apply a uniform magnetic field, then the symmetry of the Hamiltonian is broken, and there is a magnetization in the direction of the field. Figure 11.5(*a*) shows the variation of M with H at a fixed low temperature, for a large but finite system. It is indeed zero at $H = 0$, but increases rapidly when a small field is applied. As the size of the system is taken to infinity, the slope at $H = 0$ increases, and eventually becomes a discontinuity as shown in figure 11.5(*b*). In the limit, the value of M at $H = 0$ is not well defined, but the limit of M as $H \to 0$ from above or below is $\pm M_{\rm S}(T)$. I cannot reproduce here the detailed calculations that support this picture. Interested readers may like to consult, for example, Reichl (1998) or Goldenfeld (1992) for some further explanation and references to the original literature. In fact, even an infinite Ising model does not always show a spontaneous magnetization. Whether it does or not depends on the number of spatial dimensions d. It is possible to obtain an exact solution only in one and two dimensions. (By 'solution' is meant a method of calculating actual values for thermodynamic quantities like the magnetization, susceptibility and specific heat.) In one dimension, there is no ferromagnetic state. For two dimensions, the solution was given for the case of zero magnetic field in a celebrated paper by L Onsager (1944) and for a non-zero field by C N Yang (1952) and there is a spontaneous magnetization at low temperatures. For three dimensions, no exact solution has been found, but all approximate calculations indicate that there is a ferromagnetic phase. There is, in fact, very little doubt that the ferromagnetic state exists for all $d \geq 2$.

Real ferromagnets do, of course, exhibit a spontaneous magnetization and, though they may be very large, they certainly are not infinite. To understand this, we must think back to our discussion of ergodic theory. There we saw that the ensemble averages of equilibrium statistical mechanics correspond to long-time averages of the instantaneous states of an experimentally observed system. In the case of a uniaxial ferromagnet, the low-energy states with positive magnetization and those with negative magnetization constitute two separate regions of phase space. Either there are very few trajectories which can connect these two regions without passing through states of much higher energy or there are no such trajectories at all. In a large system below its Curie temperature, fluctuations in energy large enough to surmount the energy barrier between the two regions will be sufficiently rare that only one region is explored during the finite time over which the system is observed. (Just how long we would have to wait for a suitable fluctuation to occur is hard to estimate in a definitive manner, but estimates greater than the present age of the universe are sometimes quoted for systems of everyday size!) Thus, the occurrence of a spontaneous magnetization indicates a partial breakdown of ergodicity. In effect, what should be compared with observations is not the complete equilibrium ensemble average, but rather an average over half of the configurations, namely those that have a net magnetization in the same direction. This is achieved by the thermodynamic limit in the following way. When a magnetic field is applied, the statistical weight of the 'wrong' configurations is reduced relative to those of the 'right' ones. If we now take the infinite volume limit, the 'wrong' configurations are suppressed entirely, so, if we subsequently remove the field, a spontaneous magnetization remains.

Spontaneous symmetry breaking may be defined as a situation in which the Hamiltonian of a system possesses a symmetry, but the equilibrium state does not have the same symmetry. Ferromagnetism is obviously a case in point. The effective Hamiltonian (11.24) in the field-theoretic approximation to the Ising model inherits the same symmetry, $\phi \leftrightarrow -\phi$, when h is zero, and we shall shortly see how this symmetry can be spontaneously broken. Evidently, the same phenomenon must be possible in genuine relativistic field theories. For zero-temperature field theories of the kind discussed in chapter 9, the analogue of two (or more) possible states of magnetization is the existence of several possible vacuum states, one of which has been spontaneously chosen by our universe. As we shall see in chapter 12, this symmetry breaking may be invoked to explain the different strengths of the fundamental interactions. Alternatively, it could be that the universe, like a ferromagnet, possesses many domains in which the symmetry is broken in different ways.

11.5 The Ginzburg–Landau Theory

The field-theoretic approximation (11.23) and (11.24) to the Ising model is similar to the self-interacting scalar field theory we studied in chapter 9. As we saw, it is not possible to compute exactly the expectation value of any function of the field, and some further approximations must be made. One useful approximation is obtained when we evaluate the integral (11.23) by the method of steepest descent. In its simplest form, this means finding the value of ϕ at which the integrand $\exp[-H_{\mathrm{eff}}(\phi)]$ has its maximum value and replacing the integral by a constant times this maximum value of the integrand. The maximum value of the integrand corresponds to a minimum value of H_{eff}, so we have

$$Z(T, H) \approx \text{constant} \times \mathrm{e}^{-F(T,H)} \tag{11.25}$$

where $F(T, H)$ is the minimum value of H_{eff}. Apart from a constant and a factor of $1/k_{\mathrm{B}}T$ which, since we are considering only the critical region, can be replaced with $1/k_{\mathrm{B}}T_{\mathrm{c}}$, $F(T, H)$ is our approximation to the free energy. The value $M(T, H)$ of ϕ that minimizes H_{eff} is our approximation to the magnetization. This approximation constitutes the *Ginzburg–Landau* theory of phase transitions, although Ginzburg and Landau did not arrive at it in quite this way.

In general, we can allow M to depend also on position $\boldsymbol{x}$. To be a minimum of H_{eff}, it must satisfy what amounts to an Euler–Lagrange equation

$$-\nabla^2 M(\boldsymbol{x}) + r_0 M(\boldsymbol{x}) + \tfrac{1}{6}u_0 M^3(\boldsymbol{x}) = h(\boldsymbol{x}). \tag{11.26}$$

When h is independent of position, it is not hard to see that the minimum of H_{eff} occurs at the position-independent value of M which minimizes the potential

$$V(\phi) = \tfrac{1}{2}r_0\phi^2 + \tfrac{1}{4!}u_0\phi^4 - h\phi. \tag{11.27}$$

According to our earlier discussion, r_0 is positive if $T > T_0$ and negative if $T < T_0$, and $V(\phi)$ is sketched for these two cases in figure 11.6. In the high-temperature case, there is a single minimum, which is at $M = 0$ when $h = 0$. In the low-temperature case, there are two minima (if h is not too large). When $h = 0$, these two minima are at the same depth; otherwise, one or other of them is lower, according to the sign of h. This evidently corresponds, at least qualitatively, to the behaviour of a ferromagnet, if we identify T_0 as the critical temperature in this approximation. It is a simple matter to find the value of the critical exponent β in (11.9). When $h = 0$ and r_0 is negative, the solution of (11.26) for the spontaneous magnetization is

$$M_{\mathrm{S}} = \left(-\frac{6r_0}{u_0}\right)^{1/2} \propto (T_0 - T)^{1/2} \tag{11.28}$$

so we have $\beta = \frac{1}{2}$.

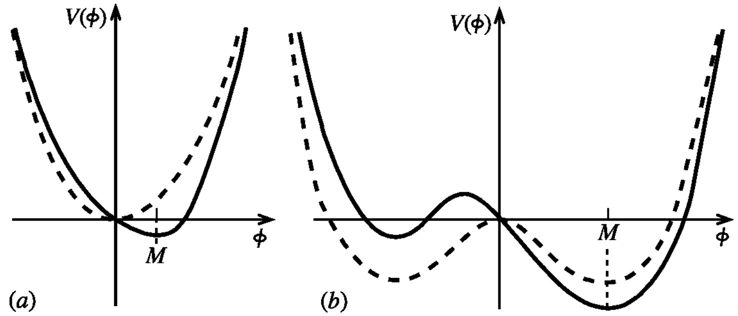


Figure 11.6. The Ginzburg–Landau potential (*a*) for $r_0 > 0$ and (*b*) for $r_0 < 0$. The symmetrical curves (broken) are for $h = 0$ and the asymmetrical ones (full) for $h > 0$.

The correlation function may be defined by analogy with (11.20), using the functional derivative discussed in appendix A, as

$$G(\boldsymbol{x}-\boldsymbol{y}) = \delta M(\boldsymbol{x})/\delta h(\boldsymbol{y}). \tag{11.29}$$

By differentiating (11.26), we find that it satisfies the equation

$$\left(-\nabla^2 + r_0 + \tfrac{1}{2}u_0 M^2\right) G(\boldsymbol{x}-\boldsymbol{y}) = \delta(\boldsymbol{x}-\boldsymbol{y}) \tag{11.30}$$

which is, not too surprisingly, the Euclidean version of (9.37) for the propagator of a scalar field. When the magnetic field h and the magnetization M are independent of position, the solution, analogous to (9.40), is

$$G(\boldsymbol{x}-\boldsymbol{y}) = \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{\exp[\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{y})]}{(k^2+m^2)} \tag{11.31}$$

where $m^2 = r_0 + \frac{1}{2}u_0 M^2$. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are far apart, this gives

$$G(\boldsymbol{x}-\boldsymbol{y}) \sim \exp(-m|\boldsymbol{x}-\boldsymbol{y}|) \tag{11.32}$$

so we identify the correlation length as

$$\xi = 1/m. \tag{11.33}$$

When $h = 0$, we have $m^2 = r_0$ above the critical temperature or, using (11.28), $m^2 = -2r_0$ below the critical temperature, and so the critical exponent for the correlation length is $\nu = \frac{1}{2}$. The susceptibility is given (up to a constant factor) by

$$\chi = \partial M/\partial h = \int \mathrm{d}^d x\, G(\boldsymbol{x}-\boldsymbol{y}) = 1/m^2 \tag{11.34}$$

and so its critical exponent is $\gamma = 2\nu = 1$.

We see that the Ginzburg–Landau theory does indeed predict critical exponents that are universal: they do not depend, for example, on u_0 or on the constant that relates r_0 to $T - T_0$, whose values vary from one magnetic system to another. It might be thought that this is an artificial result, arising from the quite drastic approximations we used to get from a real magnet or fluid to the effective Hamiltonian (11.24). This is not so, however. We could systematically improve upon these approximations by adding higher powers of ϕ and higher derivatives, and by taking more accurate account of the temperature- and field-dependence of coefficients such as r_0 and u_0. By expanding everything in powers of $T - T_0$, readers may easily convince themselves that, when M, $T - T_0$ and h are sufficiently small, all the additional terms become negligible compared with those we have retained. The only proviso is that u_0 should remain positive. If u_0 becomes zero or negative then, in order for the potential to have a minimum, a higher power of ϕ with a positive coefficient must be added, and new types of critical behaviour result (see, for example, Lawrie and Sarbach (1984)).

The critical exponents of the Ginzburg–Landau theory are the same as those obtained from a variety of simple approximations known collectively as *classical* or *mean field* theories. Other examples of such approximations are the van-der-Waals theory of imperfect gases and the Weiss molecular field theory of ferromagnetism. The reason for this is that, in all such approximations, the appropriate free energy can be written in the Ginzburg–Landau form when we are close enough to the critical point. Although the classical exponents are universal, they are only in very rough agreement with the typical experimental values I quoted earlier on. The fault lies not with the idealized model defined by (11.23) and (11.24), but with the approximation we used to estimate the functional integral. Numerous methods are available for improving on this approximation. We can, for example, return to the original Ising model (11.18) and attempt to evaluate its thermodynamic properties directly. One method of approximation is the *high-temperature series expansion* in powers of β. Since this is most accurate at very high temperatures, careful methods of extrapolation are needed to obtain results valid at the critical temperature, but good agreement with experimental values can be obtained. Another approach is the *Monte Carlo* method, which carries out the configurational sums directly by generating a set of configurations with the correct statistical weight, which should be representative of the whole ensemble. In the next section, I shall discuss an alternative approach, called the *renormalization group*, which yields rather more insight and further illustrates the analogy with relativistic quantum field theory.

11.6 The Renormalization Group

We have seen that the distinctive behaviour of a system near a critical point derives from the fact that the correlation length ξ becomes very large or, in the ideal case of an infinite system whose temperature can be adjusted to be exactly T_c, infinite.

A somewhat highbrow way of expressing this is to say that the system becomes *scale invariant* at the critical point. To see what this means, consider first the case of a finite correlation length and, to be specific, a magnet. If we examine a part of the system whose diameter is much smaller than ξ, we find that fluctuations in all the spins are strongly correlated. If, on the other hand, we examine a region whose diameter is much larger than ξ, we find that there are strong correlations only within what we now count as small subregions of diameter ξ, but not over the whole region. Thus, the appearance of the system depends upon the *length scale*, or characteristic size of the region we choose to examine. By contrast, when ξ is infinite, the appearance of the system is much the same, at whatever length scale we choose to examine it. It turns out that much valuable information about critical phenomena, including improved approximations to the critical exponents, can be obtained by investigating how the appearance of the system changes with the scale of length on which we examine it. That this might be possible was first suggested by L P Kadanoff, and detailed techniques for putting the idea into practice have been developed by many others, notably by K G Wilson and M E Fisher.

These techniques, known collectively as *renormalization-group* techniques, exist nowadays in many varied forms. Some of them are described, for example, in the books by Amit (1984), Domb and Green (1976) and Goldenfeld (1992). Here, I shall discuss one particular method, which is well suited to field-theoretic models like (11.23). In chapter 9, we found that interacting relativistic field theories require renormalization, because parameters such as masses and coupling constants appearing in the action that defines the theory do not correspond directly to measurable quantities. Here, the situation is quite similar. The parameters r_0 and u_0 in (11.24) are related to forces which act at a microscopic level, and are not best suited for describing the large-scale phenomena associated with critical points. In quantum electrodynamics, we saw that the net effect of an electric charge on, for example, the collisions of charged particles varies with the energy of the collision. This could be expressed in terms of a modified Coulomb potential or, as in (9.92), of an energy-dependent charge. Since the energy of a virtual photon exchanged in the collision can be expressed in terms of its wavelength, we can regard the energy dependence of the electric charge as a dependence on a characteristic length scale of the collision process. Furthermore, according to (9.93), the energy dependence can be related to the dependence on an arbitrary 'mass' parameter μ which may be introduced in the renormalization process. Indeed, the earliest version of the renormalization group was invented by M Gell-Mann and F E Low in just this context.

The Ginzburg–Landau theory is more or less equivalent to the lowest order of perturbation theory (for which, see chapter 9), which involves no closed-loop diagrams with momentum integrals. To obtain improved approximations for the critical exponents, it is necessary to consider higher-order contributions. Readers will recall that the momentum integrals contained in these higher-order contributions are frequently infinite, but that the infinities disappear (at least in the case of a renormalizable theory) when the results are expressed in terms of renormalized measurable quantities. In chapter 9, it appeared that these infinities were an embarrassment. Here, as we shall see, they actually work to our advantage. I shall describe only one particular calculation, that of the susceptibility exponent γ, but this will be sufficient to expose the principles that are involved. As given in (11.34), the susceptibility is the integral over all space of the correlation function, which is the Fourier transform of this function evaluated at $\boldsymbol{k} = 0$. It is actually convenient to deal with the inverse of the susceptibility, which I shall denote by $\Gamma = \chi^{-1}$. At the first order of perturbation theory, it is given by an obvious modification of equations (9.67) and (9.76), with $p = 0$:

$$\Gamma = r_0 + \tfrac{1}{2}u_0 \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{(k^2 + r_0)}. \tag{11.35}$$

The dummy integration variable k is not, of course, the k that we just set to zero.

As it stands, the integral in (11.35) is infinite if $d \geq 2$, and this infinity arises from the upper limit $|\boldsymbol{k}| \to \infty$. However, if our model field theory is regarded as an approximation to a condensed matter system such as a magnet or a fluid, then infinite values of $|\boldsymbol{k}|$ are not really allowed. For a magnet or lattice gas, the field ϕ existed originally only at the sites of a regular lattice, and $\boldsymbol{k}$ should take values only within the first Brillouin zone of the lattice. More generally, it does not make sense for a magnetization density or fluid density to vary with position over distances shorter than an atomic size, say a, so its Fourier transform has no components with $|\boldsymbol{k}| > a^{-1}$. For our purposes, it is adequate to assume that $\boldsymbol{k}$ takes values within a sphere of radius Λ, with $\Lambda \sim a^{-1}$. The integrand in (11.35) depends only on the magnitude of $\boldsymbol{k}$, so angular integrations can be carried out as in (9.76), leaving

$$\Gamma = r_0 + \tfrac{1}{2}u_0 S_d \int_0^\Lambda \mathrm{d}k \frac{k^{d-1}}{(k^2 + r_0)}. \tag{11.36}$$

The factor S_d is $(2\pi)^{-d}$ times the surface area of a unit sphere in d dimensions, which is given in appendix A. At the critical temperature, the inverse susceptibility should be zero, and we see that this now occurs when r_0 takes a value $r_{0\mathrm{c}}$ which is of order u_0. Up to corrections of order u_0^2, we find that

$$r_{0\mathrm{c}} = -\tfrac{1}{2}u_0 S_d \int_0^\Lambda \mathrm{d}k\, k^{d-3}. \tag{11.37}$$

If we define a new variable

$$t_0 = r_0 - r_{0\mathrm{c}} \tag{11.38}$$

which is proportional to $T - T_\mathrm{c}$, then (11.36) can be rewritten, again up to corrections of order u_0^2, as

$$\Gamma = t_0 \left[1 - \tfrac{1}{2}u_0 S_d \int_0^\Lambda \mathrm{d}k \, \frac{k^{d-1}}{k^2(k^2 + t_0)} \right]. \tag{11.39}$$

This certainly vanishes when $t_0 = 0$, but we want to know how it behaves when t_0 is small. The answer depends crucially on the number of spatial dimensions d. Mathematically, it is perfectly possible to take $d > 4$. In that case, the integral in (11.39) approaches a constant value at $t_0 = 0$. For small t_0, we then find that Γ is approximately a constant times t_0. Since Γ is χ^{-1}, this means that $\gamma = 1$, which is the classical value given by the Ginzburg–Landau theory. Indeed, further arguments along these lines show that in more than four dimensions, all the critical exponents of the Ginzburg–Landau theory should be exactly correct. For practical purposes, we are, of course, more interested in dimensions smaller than four. Below four dimensions, the integral in (11.39) is infinite when $t_0 = 0$, but now the infinity comes from the limit $k \to 0$. This is called an *infrared divergence* and, unlike the *ultraviolet* divergences at infinite values of k, it has a genuine physical significance, being associated with the singular behaviour of thermodynamic quantities at the critical point. To deal with the infrared divergence, we may rescale k by a factor of $t_0^{1/2}$, which gives

$$\Gamma = t_0 \left[1 - \tfrac{1}{2} u_0 t_0^{(d-4)/2} S_d \int_0^{\Lambda/t_0^{1/2}} \mathrm{d}k \, \frac{k^{d-1}}{k^2(k^2+1)} \right]. \qquad (11.40)$$

In the limit $t_0 \to 0$, the upper limit $\Lambda t_0^{-1/2}$ becomes infinite, but the integral is finite if $d < 4$. However, the factor $u_0 t_0^{(d-4)/2}$ now becomes infinite. In terms of the dimensional analysis introduced in §9.6, this quantity is dimensionless. Thus, if the expansion in (11.40) were continued to higher orders in u_0, successive terms would be proportional to successively higher powers of $u_0 t_0^{(d-4)/2}$, each term becoming infinite more rapidly than the previous one as $t_0 \to 0$.

From this it is clear that perturbation theory (the expansion in powers of u_0) does not give us a sensible answer for the dependence of the susceptibility on temperature near the critical point. The role of the renormalization group will be to reformulate perturbation theory in such a way that a sensible answer emerges. This can be done in several ways. The principle of the method I am going to explain was put forward by Wilson and Fisher (1972). The crucial observation is that expressions like (11.40) can be evaluated, in principle, when d has any value, not necessarily an integer. As a purely mathematical device, therefore, we can consider d to be a continuous real variable. The value $d = 4$ clearly marks a borderline between different kinds of critical behaviour, and it will be convenient to define a variable ϵ by

$$\epsilon = 4 - d. \qquad (11.41)$$

If we assume that the variation of the susceptibility with temperature can indeed be described by an exponent γ, then this exponent is likely to depend on ϵ. Since it is equal to 1 for any negative value of ϵ, we may anticipate that for positive values of ϵ, it can be expressed as a power series

$$\gamma = 1 + \gamma_1 \epsilon + \gamma_2 \epsilon^2 + \dots . \qquad (11.42)$$

If we can evaluate a few terms of this expansion then, by setting ϵ equal to 1, we obtain an estimate for the value of γ in three dimensions. The reason why this works is that, the smaller ϵ is, the less rapidly $u_0 t_0^{-\epsilon/2}$ diverges and the easier it becomes to extract sensible answers from perturbation theory. Clearly, any answer we obtain must be valid right up to $d = 4$ or $\epsilon = 0$. In this limit, however, the ultraviolet divergence of integrals like that in (11.40) reappears when $t_0 = 0$. The key to calculating γ is that these divergences can be removed, as we saw in chapter 9, by the process of renormalization. It should not now be too surprising that this process actually yields all the information we need to calculate γ.

As I described it in chapter 9, the object of renormalization was to express quantities like scattering amplitudes in terms of physically measurable masses and coupling constants. For our present purposes, the main object is to remove the ultraviolet divergences, and there are many different ways in which this can be achieved. Details may be found, for example, in the book by Amit (1984) and I shall just quote the results of one method. Since we have to deal with the limit $\Lambda t_0^{-1/2} \to \infty$, we might as well take Λ to be infinite at the outset. The ultraviolet divergences now appear as powers of ϵ^{-1}. They can all be removed if we express thermodynamic quantities in terms of renormalized variables u and t which, at the first order of perturbation theory, are related to u_0 and t_0 by

$$u_0 = \mu^\epsilon u \left(1 + \frac{3}{2\epsilon} S_4 u + \dots\right) \tag{11.43}$$

$$t_0 = t \left(1 + \frac{1}{2\epsilon} S_4 u + \dots\right). \tag{11.44}$$

The factor μ^ϵ in (11.43) makes the renormalized coupling constant u dimensionless. As we discussed earlier, μ is an arbitrary parameter, and u and t are variables appropriate for describing phenomena on a length scale μ^{-1}. The inverse susceptibility can now be written as

$$\Gamma = t\left[1 + \tfrac{1}{4} S_4 u \ln(t/\mu^2) + \dots\right] \tag{11.45}$$

where, to simplify matters, I have expanded in powers of ϵ as well as u and kept only the leading term. At higher orders, a wavefunction renormalization as in (9.70) also becomes necessary. Since critical phenomena are associated with very large length scales, we shall want μ to have a very small value. The way in which u and t vary with our choice of μ is expressed by differentiating (11.43) and (11.44), keeping u_0 and t_0 fixed. This leads to two functions $\beta(u)$ and $\tau(u)$, defined by

$$\beta(u) = \mu \left(\frac{\partial u}{\partial \mu}\right)_{u_0, t_0} = -\epsilon u + \tfrac{3}{2} S_4 u^2 + \dots \tag{11.46}$$

$$\tau(u) = \frac{\mu}{t}\left(\frac{\partial t}{\partial \mu}\right)_{u_0, t_0} = \tfrac{1}{2} S_4 u + \dots. \tag{11.47}$$

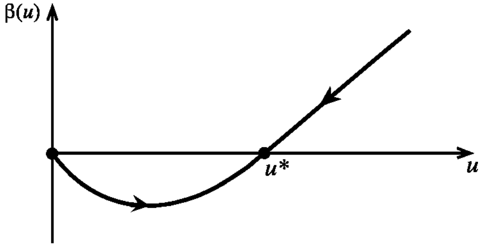


Figure 11.7. The renormalization-group function $\beta(u)$. Arrows indicate the evolution of the running coupling constant as $\mu \to 0$.

The function $\beta(u)$ is sketched in figure 11.7. It vanishes at two values of u, called *fixed points*, namely $u = 0$ and $u = u^*$, where

$$S_4 u^* = \tfrac{2}{3}\epsilon + \mathrm{O}(\epsilon^2). \tag{11.48}$$

Because $\beta(u)$ is positive for $u > u^*$ and negative for $u < u^*$, a little thought will show that u approaches the value u^* as $\mu \to 0$. In the renormalization-group approach, this is the explanation of universality. Whatever the value of u_0, which is determined in principle by the nature of microscopic forces, the renormalized coupling constant appropriate to very-large-scale phenomena is u^*.

Since u^* is of order ϵ, perturbation theory (which yields an expansion in powers of u^*) can be used to calculate the coefficients in (11.42). Suppose, indeed, that we choose $\mu^2 = t$, so that $\mu \to 0$ at the critical point. Then the inverse susceptibility (11.45) becomes just $\Gamma = t$. This might seem to imply that $\gamma = 1$, but in fact it does not, because t is not proportional to $T - T_c$. If we choose a fixed value of μ, then the corresponding renormalized coupling constant u defined by (11.43) is independent of temperature, and t, according to (11.44) is proportional to t_0 and hence to $T - T_c$. However, by choosing $\mu^2 = t$, we make u a function of t and then t is not simply proportional to t_0. To get round this, let us choose a fixed value of μ, say $\mu = \hat{\mu}$, which is sufficiently small that u can be set equal to u^* with negligible error, and let $\hat{t}$ be the corresponding renormalized temperature variable. Then $\hat{t}$ *is* proportional to $T - T_c$. For a different choice of μ, which is also small enough for u to be equal to u^*, we can relate t to $\hat{t}$ by solving the equation

$$\mu \frac{\partial t}{\partial \mu} = \tau^* t \tag{11.49}$$

where

$$\tau^* = \tau(u^*) = \tfrac{1}{3}\epsilon + \mathrm{O}(\epsilon^2) \tag{11.50}$$

with the boundary condition that $t = \hat{t}$ when $\mu = \hat{\mu}$. We get

$$t = \hat{t}(\mu/\hat{\mu})^{\tau^*}. \tag{11.51}$$

If we now set $\mu = t^{1/2}$ (which still implies $\Gamma = t$), we find

$$t \propto \hat{t}^{2/(2-\tau^*)} \propto (T - T_c)^{2/(2-\tau^*)}. \tag{11.52}$$

Finally, then, we can identify the susceptibility exponent γ as

$$\gamma = 2/(2 - \tau^*) = 1 + \tfrac{1}{6}\epsilon + O(\epsilon^2). \tag{11.53}$$

When $\epsilon = 1$, this approximation gives $\gamma = 1.17$, which is certainly an improvement on the classical value of 1. The best available estimates from more extended calculations give a value of about 1.24, in good agreement with other theoretical methods and with observations. The calculation I have described is slightly imprecise at the point where we set $u = u^*$ 'with negligible error', but it gives the correct answer, because we can take $\hat{\mu}$ as small as we like. More general and elegant (but also more long-winded) routes to the same answer can be found in several of the books mentioned in the bibliography.

More important, perhaps, than the actual values of critical exponents is the insight the renormalization group provides as to how these universal values come about. We have seen that, although they do not depend on the detailed constitution of the system, as reflected, for example, in the value of u_0, they do depend on the number of spatial dimensions d. As it turns out, they also depend on some other general features. We might, for example, generalize our field-theoretic model by taking ϕ to be an n-component vector. For $n = 3$, this would correspond to an isotropic, rather than a uniaxial, ferromagnet. It is found that the critical exponents then vary slightly with n. The susceptibility exponent, for example, is $\gamma = 1 + (n+2)\epsilon/2(n+8) + O(\epsilon^2)$, which gives a value of 1.23 when $n = d = 3$.

11.7 The Ginzburg–Landau Theory of Superconductors

The phenomena of superconductivity are both theoretically interesting and of great technological importance. For want of space, I cannot describe them in anything like the detail they deserve, and I propose mainly to highlight some theoretical considerations that turn out to have implications beyond the science of superconductivity itself. From a microscopic point of view, superconductivity is a kind of Bose–Einstein condensation. The electrons that conduct electric currents in a metal are, of course, fermions, and a non-interacting gas of fermions cannot undergo condensation. The essence of the microscopic theory is that interactions between electrons and the positive ions which form a crystal lattice can result in a net weak attraction between electrons. By analogy with quantum electrodynamics, this force can be thought of as mediated by the exchange of *phonons*, which are quantized vibrations of the lattice, much as photons are quantized 'vibrations' of the electromagnetic field. Under the influence of this attraction, some electrons may form loosely bound pairs, known as *Cooper pairs*, whose net spin is zero and which behave as bosons. These boson pairs can then undergo condensation, and the condensed electrons can flow without friction, which means that their electrical resistance is zero. A simple experimental observation that supports this picture is that the metals which superconduct most readily tend to be rather poor conductors in the normal, high-temperature state. This is because the interactions with the lattice which favour the formation of Cooper pairs cause relatively strong scattering in the normal state, which leads to a relatively high resistance. One reason for treating this qualitative picture with caution is that the mean separation of two electrons in a Cooper pair can be estimated, and it turns out to be comparable with, or greater than, the mean separation of the pairs themselves. Straightforward accounts of the microscopic theory, due originally to J Bardeen, L N Cooper and J R Schrieffer, can be found in Reichl (1998) and Tinkham (1996).

11.7.1 Spontaneous breaking of continuous symmetries

In the Ginzburg–Landau theory, the phase transition which marks the onset of superconductivity can be investigated in terms of an effective Hamiltonian similar to (11.24), in which ϕ is taken to be the macroscopic wavefunction of the condensate. This is a complex quantity, which can be expressed as

$$\phi(\boldsymbol{x}) = \frac{1}{\sqrt{2}}[\phi_1(\boldsymbol{x}) + \mathrm{i}\phi_2(\boldsymbol{x})] \qquad \text{or} \qquad \phi(\boldsymbol{x}) = \psi(\boldsymbol{x})\mathrm{e}^{\mathrm{i}\alpha(\boldsymbol{x})}. \tag{11.54}$$

The effective Hamiltonian must be real, and therefore of the form

$$H_{\mathrm{eff}}(\phi) = \int \mathrm{d}^d x \left[\nabla\phi^* \cdot \nabla\phi + r_0\phi^*\phi + \tfrac{1}{4}u_0(\phi^*\phi)^2\right]. \tag{11.55}$$

I have not included a symmetry-breaking field h, because no such field exists physically, and I have chosen the normalization of the coefficients to coincide with those of the complex scalar field in chapter 9. Whereas (11.24) has a *discrete* symmetry, $\phi \leftrightarrow -\phi$ when $h = 0$, the effective Hamiltonian (11.55) has a *continuous* symmetry, in the sense that it is unchanged if we change the phase of ϕ by any constant angle θ. This is, in fact, a gauge symmetry of the kind we studied in chapter 8. Below the critical temperature, therefore, there are not just two possible minima but an entire circle of them, as sketched in figure 11.8. Any function of the form

$$M = \langle\phi\rangle = v\mathrm{e}^{\mathrm{i}\alpha} \tag{11.56}$$

is a minimum if $v = (-2r_0/u_0)^{1/2}$ and α is any constant angle. Of course, M is not to be interpreted physically as a magnetization, but it plays a similar role in the theory. A quantity of this kind which, being non-zero in the ordered phase and zero in the disordered phase, serves to distinguish the two phases is called an *order parameter*.

It is interesting to examine fluctuations of ϕ about its mean value (11.56). Taking $\alpha = 0$ in (11.56), we write

$$\phi(\boldsymbol{x}) = \left[v + \frac{1}{\sqrt{2}}\chi(\boldsymbol{x})\right]\exp\left(\frac{\mathrm{i}\theta(\boldsymbol{x})}{\sqrt{2}\,v}\right) \tag{11.57}$$

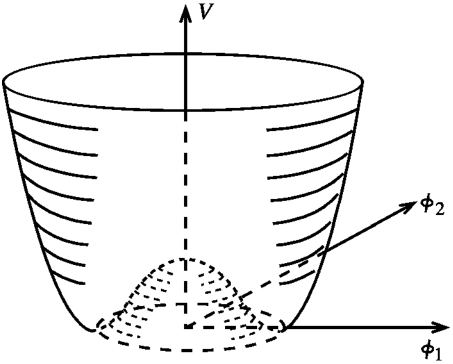


Figure 11.8. Potential for a complex scalar field with spontaneously broken symmetry. It is the surface of revolution of the symmetrical curve in figure 11.6(*b*), and its minima lie on the broken circle.

so that χ and θ measure fluctuations in the amplitude and phase, respectively, away from their mean values. Upon substituting this into the effective Hamiltonian (11.55), we obtain

$$H_{\text{eff}} = \int \mathrm{d}^d x \left[\tfrac{1}{2}\nabla\chi \cdot \nabla\chi + \tfrac{1}{2}(-2r_0)\chi^2 + \tfrac{1}{2}\nabla\theta \cdot \nabla\theta \right] + H_{\text{int}} \qquad (11.58)$$

where H_{int} contains higher powers of χ and θ, and I have dropped a constant term. If this were to be interpreted as a quantum field theory, it would represent two species of particles, the χ particles with mass $(-2r_0)^{1/2}$ and the θ particles, with zero mass, interacting through the terms in H_{int}. In the same sense that states containing such particles would be *excitations* of the vacuum state, we can speak of statistical fluctuations about the mean value of ϕ as excitations. These excitations are wave-like disturbances which, in a quantum-mechanical system, will propagate in much the same way as particles. Phonons in a solid provide an example of this. The fact that the θ excitations have zero 'mass' is easily understood from figure 11.8. A non-zero value of θ just moves ϕ around the circle of minima of the potential, which costs no potential energy. A χ fluctuation, on the other hand, moves ϕ in the radial direction, which requires an increase in potential energy. This is an example of *Goldstone's theorem*, which asserts that for any spontaneously broken continuous symmetry there is a massless particle (or 'massless' excitation), called a *Goldstone boson*.

These excitations are perhaps most easily visualized if we regard (11.55) as a model of a ferromagnet in which the spins can point with equal ease in any direction in a plane, their components being ϕ_1 and ϕ_2. The spontaneous magnetization points in one particular direction in this plane. The χ excitations are then fluctuations in the magnitude of the magnetization, while θ excitations are fluctuations in its direction. The latter are called *spin waves*, and the quantized spin waves are *magnons*. In a real ferromagnet, there are always preferred directions of magnetization, defined by the crystal lattice, and fluctuations away from these directions incur an increase in potential energy, so 'massless' magnons are not observed in practice. In superfluid helium-4, two kinds of excitations, called *phonons* and *rotons*, are found, but their detailed properties cannot be found from the condensate wavefunction alone and they do not correspond exactly to the χ and θ excitations. In superconductors, as we are about to see, the θ fluctuations have a very special effect.

11.7.2 Magnetic effects in superconductors

An important property of superconductors is the fact that they expel magnetic flux. That is, the magnetic induction $\boldsymbol{B}$ is always zero inside a superconductor. This is called the *Meissner effect*. When magnetic fields are present, the effective Hamiltonian must be modified to read

$$H_{\text{eff}} = \int \mathrm{d}^3x \left[\tfrac{1}{2}B^2 + |(\nabla - 2\mathrm{i}e\boldsymbol{A})\phi|^2 + r_0|\phi|^2 + \tfrac{1}{4}u_0|\phi|^4 - \boldsymbol{B}\cdot\boldsymbol{H} \right]. \tag{11.59}$$

The term $\frac{1}{2}B^2$ represents, in a suitable system of units, the magnetic field energy. In the next term, the gradient has been replaced by the spatial components of the covariant derivative (8.8), with $\lambda = -2$ for a Cooper pair, and the vector potential rescaled as in (8.16). In the last term, $\boldsymbol{H}$ is an externally applied magnetic field strength. In the macroscopic theory of magnetic materials, $\boldsymbol{H}$ is related to $\boldsymbol{B}$ by the equation $\boldsymbol{B} = \boldsymbol{H}+\boldsymbol{M}$, where $\boldsymbol{M}$ is the magnetization. Inside a superconductor, this implies that $\boldsymbol{M} = -\boldsymbol{H}$. It is found, though I shall not enter into the details here, that the magnetic moment of a superconducting sample is generated by a 'supercurrent' flowing on its surface. The superconductor is said to exhibit *perfect diamagnetism*. The exact relationship between the $\boldsymbol{H}$ that exists inside the superconductor and that which would be there if the sample were removed depends on the shape of the sample. For our purposes, it is sufficient to take $\boldsymbol{H}$ to be a uniform, constant field in the z direction. The magnetic induction is given in terms of the vector potential $\boldsymbol{A}$ by $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, and will undergo thermal fluctuations induced by fluctuating currents of charged particles. Thus, the partition function analogous to (11.23) includes a functional integral over $\boldsymbol{A}$ as well as ϕ; within the Ginzburg–Landau theory, H_{eff} is to be minimized with respect to both ϕ and $\boldsymbol{A}$. The term $-\boldsymbol{B}\cdot\boldsymbol{H}$ represents the energy of interaction of the magnetic moment of the superconductor with the externally applied field. Accounting properly for magnetic energy in thermodynamics is a slightly subtle matter, and is discussed (with varying degrees of clarity) in most textbooks on thermodynamics. The simplest way to see that (11.59) is correct is to consider the normal (non-superconducting) state in which $\phi = 0$. Then, by minimizing (11.59) with respect to $\boldsymbol{B}$, we find the correct result $\boldsymbol{B} = \boldsymbol{H}$.

To understand the Meissner effect, we must first find a vector potential whose curl gives a uniform magnetic induction of magnitude B in the z direction. It is easy to verify that a suitable potential is

$$\boldsymbol{A}(\boldsymbol{x}) = \tfrac{1}{2}B(-y, x, 0) \tag{11.60}$$

but other potentials, related to this one by a gauge transformation, would be equally good. Assuming that the mean value of ϕ is a constant, as in (11.56), the effective Hamiltonian becomes

$$H_{\text{eff}} = \int \mathrm{d}^3x \left[\tfrac{1}{2}B^2 + e^2B^2(x^2+y^2)|\phi|^2 + r_0|\phi|^2 + \tfrac{1}{4}u_0|\phi|^4 - BH\right]. \tag{11.61}$$

It is the second term that leads to the Meissner effect. The integral of $(x^2 + y^2)$ over the volume V of the sample is proportional to $V^{5/3}$, the exact value depending on the shape of the sample. This gives a contribution to the free energy *per unit volume* proportional to $B^2|\phi|^2V^{2/3}$, which is infinite in the thermodynamic limit, or at least very large for a macroscopic sample, if neither B nor ϕ is zero. We therefore conclude that B cannot be non-zero in a region of macroscopic size within a superconductor. There are thus two possible minima of (11.61), namely a normal state with $B = H$ and $\phi = 0$, and a superconducting state with $B = 0$ and $|\phi|^2 = -2r_0/u_0$. The free energies per unit volume of these two states are

$$F_{\text{n}}/V = -\tfrac{1}{2}H^2 \qquad \text{and} \qquad F_{\text{s}}/V = -r_0^2/u_0. \tag{11.62}$$

The stable equilibrium state is the one with the lower free energy. At a fixed temperature below T_{c}, therefore, the superconducting state is stable, provided that the applied field is smaller than a critical value given by

$$H_{\text{c}} = (2r_0^2/u_0)^{1/2}. \tag{11.63}$$

When a field larger than this is applied, the superconductivity is destroyed. Near T_{c}, the critical field varies as $H_{\text{c}} \propto (T_{\text{c}} - T)$. At lower temperatures, however, we need more detailed information about the dependence of r_0 and u_0 on T in order to find the temperature dependence of the critical field. This can be done empirically, or by deriving the effective Hamiltonian as an approximation to a detailed microscopic theory.

If we allow for mean values of B and ϕ that vary with spatial position, then other possibilities emerge, upon which I shall touch in chapter 13.

11.7.3 The Higgs mechanism

The nature of fluctuations in a superconductor is different from that envisaged in §11.7.1 because the effective Hamiltonian (11.59) is invariant under *local* gauge transformations (see chapter 8) whereas (11.55) has only a *global* gauge symmetry. Indeed, the term $\frac{1}{2}B^2 = \frac{1}{2}(\nabla \times \boldsymbol{A})^2$ in (11.59) is the three-dimensional analogue of $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ in, for example, (8.17). The magnetic induction is unchanged if we add to $\boldsymbol{A}$ the gradient of any scalar function. We can again study fluctuations by substituting (11.57) into the effective Hamiltonian (11.59). The only place where the phase fluctuation θ appears is in the covariant derivative term, which becomes

$$\frac{1}{2}\left|\nabla\chi + \mathrm{i}(\sqrt{2}v + \chi)\left(\frac{1}{\sqrt{2}v}\nabla\theta - 2e\boldsymbol{A}\right)\right|^2 . \tag{11.64}$$

Therefore, if we add to $\boldsymbol{A}$ the quantity $\nabla(\theta/2\sqrt{2}ev)$, then θ disappears entirely, and the effective Hamiltonian can be written as

$$H_{\text{eff}} = \tfrac{1}{2}\int \mathrm{d}^3x \left[|\nabla \times \boldsymbol{A}|^2 + (2\sqrt{2}ev)^2 A^2 + |\nabla\chi|^2 + (-2r_0)\chi^2\right] + H_{\text{int}} \tag{11.65}$$

where H_{int} contains higher-order terms describing self interactions of χ and interactions between χ and $\boldsymbol{A}$. We see that the excitations are χ fluctuations of 'mass' $(-2r_0)^{1/2}$ and 'photons' of mass $2\sqrt{2}ev$. In a superconductor, the 'mass' of the χ excitations is to be interpreted in terms of the correlation length $\xi = (-2r_0)^{-1/2}$, which in this context is called the *coherence length*. By analogy with (9.85), we can identify a second characteristic distance, $\lambda_{\text{p}} = 1/2\sqrt{2}ev$, called the *penetration depth*, which governs the rate of decay of magnetic forces inside a superconductor. Roughly speaking, when a magnetic field weaker than H_{c} is applied to a superconducting specimen, the magnetic induction inside the material falls off with distance x from the surface as $B(x) \sim B_0 \exp(-x/\lambda_{\text{p}})$, but the exact distribution of magnetic flux depends on the size and shape of the specimen.

It is clear that exactly the same analysis will carry over to a genuine relativistic gauge field theory. At the simplest level, we might consider the action

$$S = \int \mathrm{d}^4x \left[-\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} + (\mathrm{D}_\mu\phi)^*\mathrm{D}^\mu\phi - m_0^2\phi^*\phi - \tfrac{1}{4}\lambda_0(\phi^*\phi)^2\right] \tag{11.66}$$

where $\mathrm{D}_\mu = \partial_\mu + \mathrm{i}eA_\mu$ is the gauge-covariant derivative. When m_0^2 is positive, this describes scalar particles of charge e and their antiparticles of charge $-e$ interacting with massless photons. Bearing in mind that the photon has only two independent spin-polarization states, this gives a total of four physical degrees of freedom. When m_0^2 is negative, the gauge symmetry is spontaneously broken. The theory then describes a single scalar χ particle interacting with a massive spin-1 particle, which is no longer recognizable as a photon. The massive spin-1 particle has three independent spin states, so there are again a total of four physical degrees of freedom. We may say that one of the scalar degrees of freedom, namely the phase angle that disappears, has combined with the redundant gauge degrees of freedom to produce the third physical polarization state of the spin-1 particle. In the context of particle physics, this is known as the *Higgs mechanism* (after P Higgs, who first described it). The Higgs mechanism affords a solution to the problem we encountered in chapter 8 of constructing a gauge-invariant theory in which the gauge quanta are massive and can be identified with observed particles such as the $W^{\pm}$ and Z^0. This was the last barrier in the way of constructing a unified theory of strong, weak and electromagnetic interactions, and I shall describe this construction in the next chapter. The price to be paid is that we have to introduce scalar fields. Some of these fields, analogous to χ, should correspond to observable spin-0 particles, called *Higgs bosons*. The masses of these particles cannot be reliably predicted. At the time of writing, no such particles have been unambiguously identified by experimenters, but candidates have very recently been reported, with a mass of about 115 GeV$/c^2$.

Exercises

11.1. For a ferromagnet at its critical temperature, the magnetization is found to vary with magnetic field as $M \sim h^{1/\delta}$, where δ is a critical exponent. Show that the Ginzburg–Landau theory gives $\delta = 3$. It can often be shown that the free energy of a system near its critical point can be expressed in the *scaling* form

$$F(t, h) = |t|^{2-\alpha} f(h/|t|^{\Delta})$$

where α and Δ are two further critical exponents. Thus, up to an overall factor, it depends only on the single variable $h/|t|^{\Delta}$ rather than on h and t independently. Show that if the scaling form is correct, then the specific heat at $h = 0$ diverges as $C \sim |t|^{-\alpha}$. Show that the free energy of the Ginzburg–Landau theory does have the scaling form, with $\alpha = 0$. For any free energy that can be expressed in scaling form, show that

(a) $\beta = 2 - \alpha - \Delta$ and $\gamma = \Delta - \beta$

(b) when $y = h/|t|^{\Delta} \to \infty$, the function $f(y)$ obeys $\mathrm{d}f(y)/\mathrm{d}y \sim y^{1/\delta}$

(c) $\Delta = \beta\delta$

(d) $\gamma = \beta(\delta - 1)$

and check these results for the Ginzburg–Landau theory. The scaling property and the relations between critical exponents that follow from it are an automatic consequence of the renormalization-group analysis (see, for example, Amit (1984), Goldenfeld (1992)).

11.2. When a ferromagnet contains two or more domains, or a liquid coexists with its vapour, there is a narrow region—a domain wall or interface—between the two phases in which the magnetization or density varies quite rapidly. Consider equation (11.26) with $h = 0$ and suppose that M depends only on one spatial coordinate, say z. Show that this equation has a *soliton* solution of the form

$$M(z) = M_{\mathrm{S}} \tanh(\lambda z)$$

and identify the constant λ. Hence show that the thickness of the domain wall is approximately equal to the correlation length. Note that this applies to an *Ising* ferromagnet, in which the magnetization can point only in one of two opposite directions. In a *Bloch wall*, the magnetization *rotates* as we pass through the wall, and the thickness depends on the anisotropy energy, which is the increase in a spin's potential energy as it rotates away from the easy axis. Can you develop a variant of the Ginzburg–Landau theory to investigate this possibility? (See Lawrie and Lowe (1981).)

Chapter 12

Unified Gauge Theories of the Fundamental Interactions

We saw in chapter 8 that a special class of interacting field theories, the *gauge theories*, arise almost inevitably when we investigate the relationship between the 'internal spaces' in which fields or wavefunctions exist at different points of spacetime. We found that the simplest of these theories can be interpreted in terms of observed electromagnetic forces and, indeed, that quantum electrodynamics agrees with experimental measurements with extremely high precision. In this chapter, I shall describe how the weak and strong nuclear interactions can also be interpreted in terms of gauge theories. It would be most satisfying if the three interactions could be explained in terms not of three different gauge theories but of a single unified theory. Such theories have, as we shall see, been proposed. Just what is entailed in this unification will become clear as we proceed, but it is not entirely clear at the time of writing whether a completely unified theory can be achieved, or whether such a theory could be subjected to any very stringent experimental test.

It will, of course, be necessary to have some idea of the observed phenomena that need to be explained. High-energy particle physics is a large and rather technical subject, and it will be possible for me to give only a cursory description of the key facts that have emerged from many years of research. The weak interaction, because of its weakness, is amenable to theoretical treatment on the basis of perturbation theory and is now quite well understood. Strong-interaction phenomena, on the other hand, can often not be adequately treated by perturbation theory and, because of the difficulty of devising alternative methods of approximation, are not really understood with the same degree of confidence.

It is worth considering briefly just what 'understanding' means in this context. At the level of description that accounts for phenomena accessible to laboratory experiments, it is generally agreed that fundamental processes can be described by some kind of quantum field theory. A large part of the problem, therefore, is to be able to write down the action (or Lagrangian density) from which observed phenomena can, in principle, be derived. At the present time, it seems that an action incorporating the weak, strong and electromagnetic interactions can be written down with some confidence (though there are signs that total confidence might be misplaced). It is called the *standard model*. A second part of the problem is to be able actually to derive all the observable consequences of this model, and it is here that the strong interaction still presents difficulties. A third aspect of understanding is to decide whether the model that accounts for our current observations is truly fundamental. We have seen that large-distance or low-energy phenomena can be well described on the basis of effective Hamiltonians that bear rather little resemblance to the models we believe to represent the microscopic physical constitution of the system we study. It is entirely possible that the standard model of particle physics is itself only an effective action, valid only for the range of energies that can be produced by present-day accelerators. We shall see that there are some theoretical reasons for believing that this is indeed the case. In fact, a significant number of physicists believe that quantum field theory itself is inadequate for describing the world at a truly fundamental level, and I shall discuss some of their alternative ideas in chapter 15.

12.1 The Weak Interaction

The simplest reason for distinguishing weak, electromagnetic and strong interactions is that a hierarchy is observed in the magnitudes of quantities such as scattering cross-sections and decay rates which, on the basis of formulae such as those given in appendix D, we are inclined to attribute to a corresponding hierarchy of coupling constants. (We shall see, however, that the situation is more subtle than this.) For example, the neutral pion π^0 decays to two photons, with a mean lifetime of about 10^{-16} s. A muon, on the other hand, lives for some 10^{-6} s before decaying, through what we identify as a weak-interaction process, into an electron, a neutrino and an antineutrino. The beta decay of a free neutron into a proton, an electron and an antineutrino takes, on average, about 15 minutes, but this is exceptional even for weak interactions and is explained by the very small kinetic energies involved. The lifetimes of particles that decay by the strong interactions are typically of the order of 10^{-23} s.

In the early days of particle physics, the particles themselves were classified according to their masses into *leptons* (light particles), *baryons* (heavy particles) and *mesons* (particles of intermediate mass). In the light of improved understanding, a more detailed classification seems appropriate, which is the following. Particles which undergo strong interactions are called *hadrons*, and these can be subdivided into fermionic hadrons, the *baryons*, of which the most familiar examples are protons and neutrons, and bosonic hadrons, such as pions and kaons, which are *mesons*. Fermionic particles which have no strong interaction are called *leptons*. They include the electron, the muon and the more recently discovered tau particle, which are all negatively charged (their antiparticles being positive) and three species of neutrino, which appear to be associated with the three charged lepton species. Almost all experimental evidence is consistent with the neutrinos being exactly massless. At the time of writing, there is some indirect evidence to suggest that some or all of the neutrinos have very small, but non-zero masses. (This evidence is based on the idea of *neutrino oscillations*; the basic principle is indicated in exercise 12.1.) Whether this is really so, and, if it is, exactly how the standard model ought to be adjusted to incorporate these masses, are questions on which there is no consensus. Here, I shall describe only the simplest version of the standard model, which assumes that neutrinos have no mass. While the observed hadrons have a complicated internal structure, consistent with their being composed of more fundamental particles, the *quarks*, there is no evidence that the leptons have any internal structure. Within the standard model, the leptons are taken to be truly fundamental particles. The photon and the more recently discovered $\mathrm{W}^{\pm}$ and Z^0 particles occupy a distinguished position in this classification scheme, being (in theory) quanta of the gauge fields that mediate the electromagnetic and weak interactions. In the standard model, there are further gauge bosons, the *gluons*, associated with the strong interaction, but these, like the quarks, have not been detected in isolation.

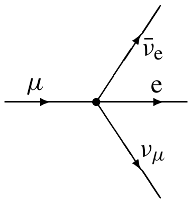
In the early 1970s, all known weak interaction phenomena could be reasonably well described by applying first-order perturbation theory to a field theory in which interactions were represented by a term in the Lagrangian density of the form

$$\mathcal{L}_{\mathrm{I}} = -\frac{1}{\sqrt{2}} G_{\mathrm{F}} \mathcal{J}_{\nu}^{\dagger}(x) \mathcal{J}^{\nu}(x). \tag{12.1}$$

An interaction of this kind, known as the *current–current* interaction, was first suggested by E Fermi. The current in question is given by

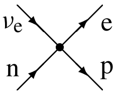
$$\mathcal{J}^{\nu}(x) = \bar{\nu}_{\mathrm{e}}(x)\gamma^{\nu}(1-\gamma^5)e(x) + \bar{\nu}_{\mu}(x)\gamma^{\nu}(1-\gamma^5)\mu(x) + \text{hadronic terms} \tag{12.2}$$

where $e(x)$ and $\mu(x)$ stand, respectively, for the electron and muon field operators, while $\nu_{\mathrm{e}}(x)$ and $\nu_{\mu}(x)$ (whose label should not be confused with a spacetime index!) are the field operators for the electron- and muon-type neutrinos. The coupling constant G_{F} is called the *Fermi constant*, and its value is given by $G_{\mathrm{F}}/(\hbar c)^3 = 1.17 \times 10^{-5}\,\mathrm{GeV}^{-2}$. The interaction (12.1) contains several terms, each giving rise to a different kind of process. For example, muon decay ($\mu^- \to \mathrm{e}^- + \bar{\nu}_{\mathrm{e}} + \nu_{\mu}$) is described by the vertex



$$-\frac{\mathrm{i}}{\sqrt{2}} G_{\mathrm{F}} \left[\bar{e}\gamma_{\nu}(1-\gamma^5)\nu_{\mathrm{e}}\right]\left[\bar{\nu}_{\mu}\gamma^{\nu}(1-\gamma^5)\mu\right] \tag{12.3}$$

where the field operator μ annihilates the decaying muon and the other three operators create the outgoing particles. The nature of the hadronic terms in (12.2) depends upon the kind of calculation we wish to undertake. For example, neutrino–neutron scattering ($\nu_{\mathrm{e}} + \mathrm{n} \to \mathrm{e}^- + \mathrm{p}$) could be described by a vertex of the form



$$-\frac{\mathrm{i}}{\sqrt{2}} G_{\mathrm{F}} \left[\bar{e}\gamma_\nu(1-\gamma^5)\nu_{\mathrm{e}}\right]\left[\bar{p}\,\Gamma^\nu n\right]. \tag{12.4}$$

In this expression, n and p are to be treated as field operators for the neutron and proton and Γ^ν is a matrix, constructed from Dirac γ matrices, which represents strong interaction effects involving the internal structure of the proton and neutron. In a theory of weak interactions only, Γ^ν is simply fitted to experimental data. In a theory that also purports to describe the strong interaction, we would instead construct contributions to the current (12.2) in terms of quark operators, of the same kind as those for the leptons. However, when we then calculate S-matrix elements as in (9.16), the 'in' and 'out' states still contain a neutron or a proton rather than free quarks, and we should have to find a means of calculating the effect of acting with quark operators on these states. This difficult task is equivalent to *calculating* Γ^ν from first principles.

The current (12.2) is called a *charged current*, because it has the net effect of raising by one unit the charge of a state on which it acts. For example, in the electronic term, $e(x)$ either annihilates a negative electron or creates a positive positron, while $\bar{\nu}_{\mathrm{e}}(x)$ creates a neutrino or annihilates an antineutrino, both of which are neutral. The form of this current and the interaction (12.1) are conjectured partly as a matter of theoretical prejudice and partly on the basis of experimental data. Since we believe the leptons to be truly fundamental particles, we expect that their interactions should be described by a simple expression, involving a minimal number of adjustable parameters. The idea of using currents is motivated by quantum electrodynamics, where the interactions of charged particles can indeed be expressed in terms of the electromagnetic current (9.79). The weak interaction currents are necessarily different, because they have to interconvert particles of different species. In principle, they might involve any or all of the bilinear covariants S, P, V^μ, A^μ and $T^{\mu\nu}$ discussed in chapter 7, with the proviso that the two field operators do not necessarily refer to the same species. The particular form that is chosen summarizes a large amount of experimental data, of which I have space only to indicate a few important features.

The most significant feature is *parity violation*. Readers will recall from chapter 7 that the parity transformation is a change of coordinates which reverses the sign of all spatial axes. This is more or less equivalent to forming the mirror image of a physical state. (Strictly speaking, account must also be taken of the *intrinsic parity* of each particle species, as is explained in any particle physics textbook, but I shall not need to make use of this.) For a long time, it was believed that parity should be a symmetry of the fundamental interactions, in the sense that any state should evolve with time in the same way as its mirror image. This means that the Lagrangian density should be unchanged by a parity transformation. It was first suggested by T D Lee and C N Yang that this symmetry is in fact violated by the weak interaction. This was confirmed experimentally by C S Wu, who studied the beta decay of cobalt-60 and found an asymmetry in the numbers of electrons emitted parallel and antiparallel to the nuclear spin. In the mirror image system, this asymmetry would be reversed, and so parity is violated. Now, each of the leptonic terms in (12.2) has the form $V^\mu - A^\mu$, where V^μ is a vector current and A^μ is an axial vector. If we consider the more general form

$$\mathcal{J}^\mu \propto (1-\alpha^2)^{1/2} V^\mu + \alpha A^\mu \tag{12.5}$$

then for the interaction we have

$$\mathcal{J}^\dagger_\mu \mathcal{J}^\mu \propto \left[(1-\alpha^2) V^\dagger_\mu V^\mu + \alpha^2 A^\dagger_\mu A^\mu\right] + \alpha(1-\alpha^2)^{1/2}\left[V^\dagger_\mu A^\mu + A^\dagger_\mu V^\mu\right]. \tag{12.6}$$

According to the transformations rules given in chapter 7, the first term is unchanged by the parity transformation, while the second changes sign. Thus, parity violation comes about through the interference between vector and axial vector currents and is a maximum when $\alpha = \pm 1/\sqrt{2}$. Thus, the $V^\mu - A^\mu$ form of the currents corresponds to *maximal parity violation*.

The reason for choosing $V^\mu - A^\mu$ rather than $V^\mu + A^\mu$ comes from the behaviour of neutrinos. We saw in chapter 7 that, for massless particles, the *chiral projections* (7.76) correspond to helicity eigenstates. Experimentally, neutrinos are always found to be emitted in the left-handed polarization state, while antineutrinos are always right-handed. Readers should be able to convince themselves that only these states can be created by the $V^\mu - A^\mu$ current interaction.

It is, of course, possible to write down more general interactions involving the S, P and $T^{\mu\nu}$ covariants. When applied to muon decay, nuclear beta decay and neutrino-nucleus scattering, the various terms lead to different dependences on the angles between momenta and spins of the various particles involved. These place quite stringent limits on any possible contributions from scalar or tensor interactions. A sensitive test for the presence of pseudoscalar interactions is provided by the decay of charged pions. These decays almost always produce a muon and a neutrino, but a fraction of about 1.27×10^{-4} of pion decays produce instead an electron and a neutrino. Calculations show that if the interaction were entirely pseudoscalar, then the electronic decays would, on the contrary, be about five times more frequent than the muonic ones. Calculations based on the $V^\mu - A^\mu$ interaction, however, agree well with the observed ratio, so any pseudoscalar interaction must be extremely small. This close agreement also provides good evidence for *electron-muon universality*, which refers to the fact that the electron and muon currents appear in (12.2) with the same weight and therefore have weak interactions of the same strength. The value of G_F can be found by comparing calculated lifetimes both of muons and of nuclei that undergo beta decay with experimentally measured values, and consistent results are obtained by these two methods.

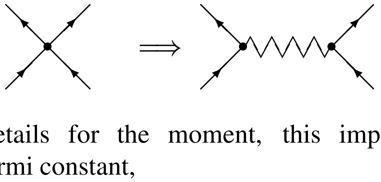
Because of the difficulty of carrying out reliable strong interaction calculations, there is less detailed information about the form of hadronic currents. If, in the vertex (12.4), it is assumed that

$$\Gamma^\mu = \gamma^\mu(C_V + C_A\gamma^5) \tag{12.7}$$

then it is found that $C_A/C_V \approx -1.26$. This can be taken as evidence for an underlying $V^\mu - A^\mu$ structure in the hadronic currents also.

Although the current–current interaction is able to account for quite a large body of observed low-energy phenomena, it has some important shortcomings. One is that there are some phenomena for which it cannot account, as we shall see. Theoretically, it has two highly undesirable features. One is that it does not satisfy the requirement of *unitarity*. Reduced to its simplest terms, this requirement means that, given an initial state, the total probability of observing *some* final state must be 1. More technically, it means that the scattering operator S, which transforms 'out' states into 'in' states as in (9.5), must be unitary. From this is can be shown to follow that the total cross-section for, say, electron–neutrino scattering must decrease at high energies at least as fast as constant/q^2, where q is the total 4-momentum. When such cross-sections are calculated from the Fermi theory, they are found to *increase* as $G_{\mathrm{F}}^2 q^2$, as might be expected from dimensional analysis, so unitarity is violated. A related problem is that the theory is not renormalizable. Since the coupling constant G_{F} has the dimensions (energy)$^{-2}$, the dimensional criterion for renormalizability discussed in chapter 9 is not satisfied. At all orders of perturbation theory beyond the first, there are infinities that cannot be renormalized away and the theory does not make sense.

The accepted cure for these problems is to introduce an *intermediate vector boson*. If the field operator for this spin-1 particle is W^μ, then the current–current interaction is replaced by something like $g(\mathcal{J}_\mu^\dagger W^\mu + \mathcal{J}_\mu W^{\mu\dagger})$, since the action must be Hermitian. This is obviously similar to the electromagnetic interaction in (9.78) and, in particular, the new coupling constant g is dimensionless. The effect of this replacement upon processes of the kind we have been considering is to split a four-fermion vertex like those in (12.3) and (12.4) into a pair of vertices of the kind that occur in QED, connected by a W propagator:



Ignoring technical details for the moment, this implies a corresponding replacement for the Fermi constant,

$$G_{\mathrm{F}} \Longrightarrow \frac{-g^2}{k^2 - M_{\mathrm{W}}^2} \tag{12.8}$$

where k is the 4-momentum transferred between the two halves of the vertex. When the magnitude of this 4-momentum is much smaller than the mass M_W of the intermediate vector boson, this is just a constant, and we recover the Fermi theory with $G_F = g^2/M_W^2$. At high energies, however, the composite vertex behaves as $-g^2/k^2$ and this, other things being equal, solves the problems of unitarity and renormalizability.

Models of the weak interaction based on this idea of an intermediate vector boson were suggested by S L Glashow (1961) and by A Salam and J C Ward (1964), but they lacked the crucial property of gauge invariance which, as we saw in chapter 9, is essential for a theory containing spin-1 particles to be renormalizable. The missing ingredient was the Higgs mechanism, discussed in the previous chapter, which allows masses for the spin-1 particles to by generated within a gauge-invariant theory by spontaneous symmetry breaking. A highly successful model that incorporates the Higgs mechanism was devised by S Weinberg (1967) and by Salam (1968). At the time, it was not entirely clear whether even this model would really be renormalizable, but its renormalizability was finally proved by G ’t Hooft (1971).

12.2 The Glashow–Weinberg–Salam Model for Leptons

The Glashow–Weinberg–Salam model (which I shall abbreviate henceforth to GWS) is a non-Abelian gauge theory. As I explained it in chapter 8, these theories involve grouping observed particles into multiplets and regarding the members of a multiplet as different states of the same basic particle. Our problem is, of course, to decide which groups of particles nature actually does regard in this way. The groupings that have been found to work involve a further subtlety, which may appear strange at first sight. It will be convenient at the beginning to imagine that both the electron and its neutrino are massless and to endow the electron with a mass at a later stage. As we saw in §7.5, the left- and right-handed components (7.76) of the field for a massless fermion can be treated quite independently. Since right-handed neutrinos are not observed, we can assume that they do not exist.

Consider now the electronic part of the current (12.2). It will be convenient to redefine it by inserting a factor of $\frac{1}{2}$. Because of the anticommutation relation $\gamma^\mu\gamma^5 = -\gamma^5\gamma^\mu$, we see that it involves only the left-handed components of both the neutrino and the electron:

$$\mathcal{J}_e^\nu = \bar{\nu}_e\gamma^\nu\tfrac{1}{2}(1-\gamma^5)e = \bar{\nu}_e\tfrac{1}{2}(1+\gamma^5)\gamma^\nu\tfrac{1}{2}(1-\gamma^5)e = \bar{\nu}_{eL}\gamma^\nu e_L. \qquad (12.9)$$

These two left-handed components are assigned to a doublet, analogous to the nucleon doublet (8.18). We write

$$\ell_e = \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} \qquad (12.10)$$

the notation indicating a doublet of left-handed electron-type particles. This commits us to an SU(2)×U(1) gauge theory like that discussed in §8.3, and the

SU(2) property is called *weak isospin* to distinguish it from the nuclear isotopic spin. The doublet has, of course, a weak isospin of $t = \frac{1}{2}$, with $t^3 = +\frac{1}{2}$ for the neutrino and $t^3 = -\frac{1}{2}$ for the electron. To get the correct electric charges from the Gell-Mann–Nishijima formula (8.54). we assign to the doublet a *weak hypercharge* of $y = -1$. As in §8.2, we now use the Pauli matrices to represent the current (12.9) as

$$\mathcal{J}_{\mathrm{e}}^{\mu} = \bar{\ell}_{\mathrm{e}}\gamma^{\mu}\tau^{+}\ell_{\mathrm{e}} \tag{12.11}$$

where

$$\tau^{+} = \tfrac{1}{2}(\tau^1 + \mathrm{i}\tau^2) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{12.12}$$

The coupling of this current to a gauge field, say W, must contribute to the Lagrangian density an Hermitian operator, consisting of the two terms $\mathcal{J}_{\mathrm{e}}^{\mu}W_{\mu} + \mathcal{J}_{\mathrm{e}}^{\mu\dagger}W_{\mu}^{\dagger}$. We can write this more explicitly, including a coupling constant g, which measures the strength of the interaction, as

$$\mathcal{L}_{\mathrm{I}} = -\frac{g}{\sqrt{2}}\bar{\ell}_{\mathrm{e}}(\tau^{+}\not{W}^{+} + \tau^{-}\not{W}^{-})\ell_{\mathrm{e}} = -\frac{g}{2}\bar{\ell}_{\mathrm{e}}(\tau^{1}\not{W}_{1} + \tau^{2}\not{W}^{2})\ell_{\mathrm{e}} \tag{12.13}$$

where $\tau^{-} = \frac{1}{2}(\tau^1 - \mathrm{i}\tau^2)$ and, in the second expression,

$$W_{\mu}^{1} = 2^{-1/2}(W_{\mu}^{+} + W_{\mu}^{-}) \qquad \text{and} \qquad W_{\mu}^{2} = 2^{-1/2}\mathrm{i}(W_{\mu}^{+} - W_{\mu}^{-}).$$

The current $\mathcal{J}_{\mathrm{e}}^{\mu}$ acting on any state increases the charge by one unit, either annihilating an electron or creating a positron. To conserve electric charge, the field W_{μ}^{+} must annihilate a positive gauge boson W^{+} or create its negatively charged antiparticle W^{-}; the adjoint field operator $W_{\mu}^{-} = W_{\mu}^{+\dagger}$ has the converse effect. This form of interaction will reproduce the Fermi theory of charged weak currents (so far as the electron-type particles on their own are concerned) in the manner I described qualitatively in the previous section. It will not yet, however, lead to a gauge-invariant theory. By comparison with the SU(2) theory developed in chapter 8, we see that a third gauge field, W_{μ}^{3}, coupled to a new current, is required to make the interaction invariant under weak isospin rotations. Thus, we must enlarge (12.13) to read

$$\mathcal{L}_{\mathrm{I}} = -\frac{g}{2}\bar{\ell}_{\mathrm{e}}(\tau^{1}\not{W}_{1} + \tau^{2}\not{W}^{2} + \tau^{3}\not{W}^{3})\ell_{\mathrm{e}} = -g\bar{\ell}_{\mathrm{e}}\boldsymbol{t}\cdot\not{\boldsymbol{W}}\ell_{\mathrm{e}} \tag{12.14}$$

where the three matrices $\boldsymbol{t} = \frac{1}{2}\boldsymbol{\tau}$ are the generators of the isospin-$\frac{1}{2}$ representation. The new current, given by

$$\bar{\ell}_{\mathrm{e}}\tau^{3}\gamma^{\mu}\ell_{\mathrm{e}} = \bar{\nu}_{\mathrm{eL}}\gamma^{\mu}\nu_{\mathrm{eL}} - \bar{e}_{\mathrm{L}}\gamma^{\mu}e_{\mathrm{L}} \tag{12.15}$$

is a *neutral current*, which has no net effect on the charge of a state on which it acts. The second term is clearly proportional to the electromagnetic current. As we shall see, however, the gauge invariant theory also involves a *weak* neutral current and thus predicts new interaction effects, which have indeed been observed.

In order to incorporate electromagnetism correctly, it is necessary to include a fourth gauge field B_μ associated with phase transformations. As we saw in chapter 8, the U(1) group of electromagnetism is not the same as the U(1) group of phase transformations. In accordance with the Gell-Mann–Nishijima formula, we shall find that the electromagnetic field A_μ is a linear combination of B_μ and W_μ^3. It is, of course, most gratifying that this leads to a description of both weak and electromagnetic forces within a single framework, and it should be noted that we cannot treat the weak interaction in isolation by ignoring the phase transformations. (Readers should be able to satisfy themselves that this would be possible if and only if the electron and neutrino had the same charge.) At this point, the total Lagrangian density reads

$$\mathcal{L} = -\tfrac{1}{4}F_{\mu\nu}^{(W)}F^{(W)\mu\nu} - \tfrac{1}{4}F_{\mu\nu}^{(B)}F^{(B)\mu\nu} + \bar{\ell}_\mathrm{e}\gamma^\mu\left(\mathrm{i}\partial_\mu - g\boldsymbol{t}\cdot\boldsymbol{W}_\mu - g'\tfrac{1}{2}yB_\mu\right)\ell_\mathrm{e} \tag{12.16}$$

where the field strength tensor $F_{\mu\nu}^{(W)}$ is constructed from $\boldsymbol{W}_\mu$ in the same way as (8.37), with the SU(2) structure constants $C^{abc} = \epsilon^{abc}$, and $F_{\mu\nu}^{(B)}$ from B_μ as in (8.14). The two coupling constants g and g' associated with the two groups SU(2) and U(1) are independent. This Lagrangian density is invariant under the SU(2)×U(1) gauge transformations

$$\begin{aligned}
\ell_\mathrm{e} &\to \exp\left[\mathrm{i}\tfrac{1}{2}y\theta(x) + \mathrm{i}\boldsymbol{\alpha}(x)\cdot\boldsymbol{t}\right]\ell_\mathrm{e} \equiv \exp\left[\mathrm{i}\tfrac{1}{2}y\theta(x)\right]U(\boldsymbol{\alpha})\ell_\mathrm{e} \\
W_\mu &\to U(\boldsymbol{\alpha})W_\mu U^{-1}(\boldsymbol{\alpha}) + (\mathrm{i}/g)[\partial_\mu U(\boldsymbol{\alpha})]U^{-1}(\boldsymbol{\alpha}) \\
B_\mu &\to B_\mu - (1/g')\partial_\mu\theta.
\end{aligned} \tag{12.17}$$

As in chapter 8, the matrix W_μ is defined as $\boldsymbol{t}\cdot\boldsymbol{W}_\mu$.

So far, neither the gauge bosons nor the electron have masses. To put this right, without losing the gauge invariance, we must introduce a Higgs scalar field, as described in chapter 11. In the simplest version of the GWS theory, it is an SU(2) doublet

$$\phi = \begin{pmatrix}\phi^+ \\ \phi^0\end{pmatrix}. \tag{12.18}$$

The component ϕ^0 will be given a vacuum expectation value v

$$\langle 0|\phi|0\rangle = \begin{pmatrix}0 \\ v\end{pmatrix} \tag{12.19}$$

so, since the vacuum contains no electric charge, the Higgs doublet must have hypercharge $y = 1$, making the ϕ^0 particles neutral. We add to the Lagrangian density (12.16) the quantity

$$\mathcal{L}_\mathrm{Higgs} = (\mathrm{D}_\mu\phi)^\dagger(\mathrm{D}^\mu\phi) - \tfrac{1}{4}\lambda\left[(\phi^\dagger\phi) - v^2\right]^2 \tag{12.20}$$

where the covariant derivative is

$$\mathrm{D}_\mu\phi = \left[\partial_\mu + \mathrm{i}g\boldsymbol{t}\cdot\boldsymbol{W}_\mu + \mathrm{i}g'\tfrac{1}{2}B_\mu\right]\phi. \tag{12.21}$$

Mathematically, of course, any constant value of ϕ such that $\phi^\dagger\phi = v^2$ is a minimum of the potential in (12.20). By making a gauge transformation, it is always possible to bring this expectation value into the form (12.19). This transformation will also rearrange the components of the electron–neutrino doublet. Physically, the particles we recognize as electrons and neutrinos are those created and annihilated by the field operators that appear in this doublet after the transformation has been made.

To find the masses of the gauge bosons, we set

$$\phi(x) = \langle 0|\phi|0\rangle + \widetilde{\phi}(x) \tag{12.22}$$

which gives

$$\mathcal{L}_{\mathrm{Higgs}} = \tfrac{1}{2}(gv)^2 W_\mu^+ W^{-\mu} + \tfrac{1}{4}v^2(gW_\mu^3 - g'B_\mu)(gW^{3\mu} - g'B^\mu) + \ldots \tag{12.23}$$

where the terms represented by $\ldots$ are those that describe the particles created and annihilated by $\widetilde{\phi}$ and their interactions with the gauge bosons. From the first term, we identify the mass of the W$^+$ particle and its antiparticle, the W$^-$, as

$$M_{\mathrm{W}}^2 = \tfrac{1}{2}(gv)^2. \tag{12.24}$$

The second term contains a linear combination of W_μ^3 and B_μ, which is to be identified as the field operator for a third weak gauge boson, the Z^0. To make sure that this field operator creates and annihilates particle states with our standard normalization (7.18), it must be of the form

$$Z_\mu = \cos\theta_{\mathrm{W}} W_\mu^3 - \sin\theta_{\mathrm{W}} B_\mu. \tag{12.25}$$

The angle θ_{W} is called the *weak mixing angle* or the *Weinberg angle*; it is introduced simply to make sure that the squares of the two coefficients sum to 1. Its value is not known *a priori*, but it can be measured, by methods I shall mention shortly, and is found to be given by

$$\sin^2\theta_{\mathrm{W}} \approx 0.22 \qquad \text{or} \qquad \theta_{\mathrm{W}} \approx 28^\circ. \tag{12.26}$$

If the field Z_μ as defined by (12.25) is to be proportional to the combination $gW_\mu^3 - g'B_\mu$ that appears in (12.23), then we must have

$$\tan\theta_{\mathrm{W}} = g'/g. \tag{12.27}$$

In that case, the second term of (12.23) is $\frac{1}{2}M_{\mathrm{Z}}^2 Z_\mu Z^\mu$, with the mass of the Z^0 given by

$$M_{\mathrm{Z}}^2 = \frac{1}{2}\left(\frac{gv}{\cos\theta_{\mathrm{W}}}\right)^2 = \frac{M_{\mathrm{W}}^2}{\cos^2\theta_{\mathrm{W}}}. \tag{12.28}$$

The gauge field A_μ of electromagnetism is a second linear combination of W^3_μ and B_μ. It should also create and annihilate particle states with the correct normalization. Moreover, since the photon and the Z^0 are distinct particles, the creation and annihilation operators in A_μ must commute with those in Z_μ. Both of these criteria are met if we define

$$A_\mu = \cos\theta_{\mathrm{W}} B_\mu + \sin\theta_{\mathrm{W}} W^3_\mu. \tag{12.29}$$

We should check that the A_μ defined in this way really does correspond to the electromagnetic field. To do this, we consider the special gauge transformation specified by (8.50). The factor of $\frac{1}{2}y$ is already included in the gauge transformation (12.17), so we simply take $\alpha_1 = \alpha_2 = 0$ and $\alpha_3 = \theta$. For the fields defined by (12.25) and (12.29) and the fields $W^\pm_\mu$ of the charged gauge bosons, this gauge transformation gives

$$\begin{aligned} Z_\mu &\to Z_\mu \\ A_\mu &\to A_\mu - \left(\frac{\cos\theta_{\mathrm{W}}}{g'} + \frac{\sin\theta_{\mathrm{W}}}{g}\right)\partial_\mu\theta \\ W^\pm_\mu &\to \mathrm{e}^{\pm\mathrm{i}\theta} W^\pm_\mu. \end{aligned} \tag{12.30}$$

These are exactly what we expect for the electromagnetic gauge transformation, provided that the change in A_μ can be identified as $-(1/e)\partial_\mu\theta$. Together with (12.27), this tells us that the fundamental electric charge is given in terms of the SU(2) and U(1) coupling constants by

$$e = gg'/(g^2 + g'^2)^{1/2}. \tag{12.31}$$

Finally, we must arrange for the electron to have a mass. This requires a term in the Lagrangian equal to $-m\bar{e}e = -m(\bar{e}_{\mathrm{L}}e_{\mathrm{R}} + \bar{e}_{\mathrm{R}}e_{\mathrm{L}})$. In the standard version of the GWS model, the right-handed component e_{R} is treated on a separate footing from e_{L}. Since e_{R} does not appear in the weak currents, it is unaffected by the SU(2) transformations and is therefore assigned a weak isospin $t = 0$ (it is a weak-isospin *singlet*). To get its charge right, it must have a hypercharge $y = -2$. For this reason, the mass term quoted above is not gauge invariant. The electron mass can be generated in a gauge-invariant manner from spontaneous symmetry breaking. We add to $\mathcal{L}$ the gauge-invariant expression

$$\Delta\mathcal{L}_{\mathrm{e}} = \bar{e}_{\mathrm{R}}\mathrm{i}\gamma^\mu(\partial_\mu - \mathrm{i}g'B_\mu)e_{\mathrm{R}} - f_{\mathrm{e}}(\bar{\ell}_{\mathrm{e}}\phi e_{\mathrm{R}} + \bar{e}_{\mathrm{R}}\phi^\dagger\ell_{\mathrm{e}}) \tag{12.32}$$

where f_{e} is a constant. The contribution to this from the vacuum expectation value of ϕ gives the required mass term with

$$m = f_{\mathrm{e}}v. \tag{12.33}$$

The muon, the tau lepton and their associated neutrinos can now be incorporated by adding to $\mathcal{L}$ further terms of exactly the same form as those involving the electron and its neutrino.

12.3 Physical Implications of the Model for Leptons

As far as the electroweak interactions of leptons are concerned, the model is now complete. The easiest way to see its implications for physical phenomena at low energies (that is, at energies much smaller than the masses of the $\mathrm{W}^{\pm}$ and Z^0 bosons) is to derive an effective Lagrangian density with an interaction term similar to that of the Fermi theory (12.1). The particles associated with the Higgs field $\widetilde{\phi}$ have not been unambiguously identified amongst the products of scattering events (although, as I mentioned in chapter 11, a few candidates for such particles have, at the time of writing been tentatively identified) so they must have masses at least as large as those of the gauge bosons. At low energies, therefore, their propagators are small and make a negligible contribution to observed processes. We can eliminate them by setting ϕ equal to its vacuum expectation value. For processes involving energies much smaller than M_{W} and M_{Z}, the important terms in $\mathcal{L}$ that involve the weak gauge bosons and their interactions with the leptons can be written as

$$\hat{\mathcal{L}} = M_{\mathrm{W}}^2 W_\mu^+ W^{-\mu} + \tfrac{1}{2} M_{\mathrm{Z}}^2 Z_\mu Z^\mu - \frac{g}{\sqrt{2}}(W_\mu^+ \mathcal{J}^\mu + W_\mu^- \mathcal{J}^{\dagger\mu}) - \frac{g}{\cos\theta_{\mathrm{W}}} Z_\mu \mathcal{J}_0^\mu . \tag{12.34}$$

The first two terms come from the Higgs-field Lagrangian (12.20) and the others from the leptonic part of (12.16) and the gauge-field term in (12.32), together with similar terms for the other lepton species. The charged current $\mathcal{J}^\mu$ is (12.9) with additional muon and tau terms. The neutral current that couples to Z_μ is

$$\mathcal{J}_0^\mu = \tfrac{1}{2}\bar{\nu}_{\mathrm{eL}}\gamma^\mu \nu_{\mathrm{eL}} + (\sin^2\theta_{\mathrm{W}} - \tfrac{1}{2})\bar{e}_{\mathrm{L}}\gamma^\mu e_{\mathrm{L}} + \sin^2\theta_{\mathrm{W}}\bar{e}_{\mathrm{R}}\gamma^\mu e_{\mathrm{R}} + \ldots \tag{12.35}$$

again with additional muon and tau terms.

As far as W_μ and Z_μ are concerned, (12.34) is a quadratic form. Remembering that the Lagrangian density we have constructed is to be used in a functional integral, the integral over W_μ and Z_μ can be carried out in much the same way that we used, for example, to obtain the generating functional (9.41). Defining the effective Fermi interaction by

$$\int \mathcal{D}W\,\mathcal{D}Z \,\exp\left(\mathrm{i}\int \mathrm{d}^4x\,\hat{\mathcal{L}}\right) = \text{constant} \times \exp\left(\mathrm{i}\int \mathrm{d}^4x\,\mathcal{L}_{\mathrm{I,eff}}\right) \tag{12.36}$$

we find

$$\mathcal{L}_{\mathrm{I,eff}} = -\frac{g^2}{2M_{\mathrm{W}}^2}\left(\mathcal{J}_\mu^\dagger \mathcal{J}^\mu + \mathcal{J}_{0\mu}\mathcal{J}_0^\mu\right). \tag{12.37}$$

The first, charged current, term has the same form as (12.1), except that the currents in the GWS theory differ from those in the Fermi theory by a factor of $\frac{1}{2}$. We can therefore identify the Fermi constant as

$$G_{\mathrm{F}} = g^2/4\sqrt{2}M_{\mathrm{W}}^2. \tag{12.38}$$

From (12.27) and (12.31), we find that $g = e/\sin\theta_{\mathrm{W}}$, so this can be rearranged to express the W mass as

$$M_{\mathrm{W}}^2 = e^2/4\sqrt{2}G_{\mathrm{F}}\sin^2\theta_{\mathrm{W}}. \tag{12.39}$$

The values of e and G_{F} are well known from experiment, so we can now *predict* the mass of the $\mathrm{W}^{\pm}$ and, from (12.28), the mass of the Z^0, provided that the Weinberg angle can be ascertained. This angle appears in the neutral current (12.35), which is an addition to the Fermi theory. The neutral current leads to new processes such as the elastic scattering of neutrinos by electrons. The neutrino beams needed to observe these processes first became available in the early 1970s, when the predicted neutral current effects were indeed found, giving the first experimental evidence in favour of the GWS theory. The value of $\sin^2\theta_{\mathrm{W}}$ emerging from these experiments was 0.217 ± 0.014. From this value, we get the following predictions for the W and Z masses:

$$M_{\mathrm{W}} = 80.2 \pm 2.6\,\mathrm{GeV} \qquad M_{\mathrm{Z}} = 90.6 \pm 2.1\,\mathrm{GeV}. \tag{12.40}$$

When these particles were actually observed at CERN in 1982–3, their masses were found to be $M_{\mathrm{W}} = 80.8 \pm 2.7\,\mathrm{GeV}/c^2$ and $M_{\mathrm{Z}} = 92.9 \pm 1.6\,\mathrm{GeV}/c^2$, giving convincing evidence to support the GWS theory.

Since that time, the standard model has been subjected to precise tests, through studies of scattering and decay processes that are far too extensive for me to give any useful summary here. Interested readers may like to consult, for example, Donoghue *et al* (1994), Barnett *et al* (1996) and Groom *et al* (2000). A point worth emphasizing is that our discussion has taken no account of the higher-order corrections in perturbation theory which, according to §9.6, lead to renormalization of the parameters appearing in the Lagrangian density. Because of the weakness of both the weak and the electromagnetic interactions, the effects of higher-order corrections are small, but experimental precision is such that they must be taken into account. The renormalized W and Z masses correspond to quantities that can be unambiguously defined in experimental terms; the most accurate values available as I write are

$$M_{\mathrm{W}} = 80.419 \pm 0.056\,\mathrm{GeV}/c^2 \qquad M_{\mathrm{Z}} = 91.1882 \pm 0.0022\,\mathrm{GeV}/c^2. \tag{12.41}$$

The weak mixing angle θ_{W}, however, is not a directly measurable quantity and it can be defined in several different ways, which are not entirely equivalent. One way is to use (12.28) to define the renormalized θ_{W} as $\cos\theta_{\mathrm{W}} = M_{\mathrm{W}}/M_{\mathrm{Z}}$. According to this definition, it has approximately the value given in (12.26). There are, though, several reasons why an alternative definition might be preferable. One is that M_{W} has been determined less accurately than M_{Z}, and also less accurately than other parameters, such as the Fermi constant G_{F}. Another is that θ_{W} plays a rather more fundamental role in the theory as representing the ratio of the two coupling constants g and g', as shown in (12.27). An alternative definition of θ_{W} is arrived at by first defining renormalized versions of these coupling constants. A method commonly used is the so-called $\overline{\mathrm{MS}}$ renormalization procedure, which is similar, though not quite identical, to the one used in (11.43) (but taking the limit $\epsilon \to 0$ rather than $\epsilon = 1$). These are 'running' coupling constants, in the sense we discussed in §§9.7 and 11.6, and are normally evaluated with $\mu = M_{\mathrm{Z}}$. Taking $\tan\theta_{\mathrm{W}}$ as the ratio of these renormalized coupling constants, one can obtain a theoretical expression for M_{Z} of the form

$$M_{\mathrm{Z}}^2 = \frac{e^2}{4\sqrt{2}G_{\mathrm{F}}K\sin^2\theta_{\mathrm{W}}\cos^2\theta_{\mathrm{W}}} \tag{12.42}$$

which follows from (12.28) and (12.39), except for the quantity K, which is close to 1, but takes account of higher-order corrections. The weak mixing angle $\hat{\theta}_{\mathrm{W}}$ defined in this way can be determined from the measured values of M_{Z}, G_{F} and other information needed to estimate K, with the result

$$\sin^2\hat{\theta}_{\mathrm{W}} = 0.23117 \pm 0.00016. \tag{12.43}$$

With this definition of $\hat{\theta}_{\mathrm{W}}$, the relation (12.28) is an independent *prediction* of the standard model, which can be subjected to further tests. For example, it is useful to define a parameter

$$\rho = \frac{M_{\mathrm{W}}^2}{K' M_{\mathrm{Z}}^2\cos^2\hat{\theta}_{\mathrm{W}}} \tag{12.44}$$

where K' is again close to 1 but includes higher-order corrections. According to the standard version of the theory, ρ should be exactly equal to 1, but modified versions, such as the one explored in exercise 12.3 give different values. The value of ρ consistent with a variety of experimental data is

$$\rho = 1.001 \pm 0.003 \tag{12.45}$$

in good agreement with the standard version of the theory.

12.4 Hadronic Particles in the Electroweak Theory

12.4.1 Quarks

The idea that the hadrons are composed of *quarks* was first put forward by M Gell-Mann and G Zweig in the early 1960s. The species or *flavours* that are currently thought to exist, together with their electric charges Q in units of e, are

up (u)	charmed (c)	top (t)	$Q = \frac{2}{3}$
down (d)	strange (s)	bottom (b)	$Q = -\frac{1}{3}$.

None of these particles has ever been detected in isolation and, as we shall see later, they are believed to be permanently confined inside the hadrons that are

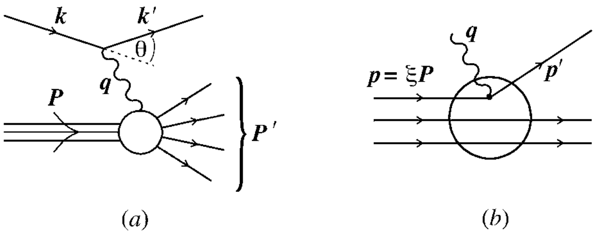


Figure 12.1. Schematic view of the deep inelastic scattering of an electron by a proton. The net effect of the collision, shown in (*a*), is to produce, in addition to the scattered electron, a collection of hadronic particles with total momentum P'. If the virtual photon has a short enough wavelength, it strikes a single charged constituent of the proton, as indicated in the close-up view (*b*), and the disrupted proton subsequently 'fragments' to form the debris indicated in (*a*).

observed. For this reason, their masses cannot be unambiguously determined. Such estimates as can be obtained suggest masses ranging from around 5 MeV$/c^2$ for the u and d quarks to some 174 GeV$/c^2$ for the t quark. Particles containing the t quark were first identified experimentally in 1995 at the Fermi National Accelerator Laboratory. The mass of this quark can be determined with fair confidence, being much larger than any other contributions to the total mass of a particle that contains it.

There are several kinds of evidence for the existence of quarks. The masses and magnetic moments of all the observed hadrons can be reasonably well accounted for by modelling them as bound states of quarks, each baryon being composed of three quarks and each meson of a quark and an antiquark. A few examples are the proton (uud), neutron (udd), Ω^- (sss), π^+ (u$\bar{\mathrm{d}}$) and K^0 (d$\bar{\mathrm{s}}$). All the particles expected on this basis are observed and all observed particles fit into the scheme. The transformations of observed particle species that occur in scattering and decay events are all consistent with rearrangements of their quark contents.

Moreover, the dependence of scattering cross-sections at high energies on energy and scattering angles is characteristic of that expected for scattering of point-like constituent particles, a fact somewhat analogous to the strong back-scattering of α particles which led Rutherford to postulate the existence of atomic nuclei. The nature of this crucial piece of evidence for the actual existence of quarks is worth understanding in a little more detail. As an example, consider the collision of a high-energy electron with a stationary proton depicted in figure 12.1(*a*)—a process known as *deep inelastic scattering*. It is a reasonable approximation to suppose that this process comes about through the mediation of a single virtual photon, because corrections due to the exchange of more photons are small, having additional factors of the fine structure constant α. At high energies, the hadronic debris emerging from the collision (whose net 4-

momentum is denoted by P' in figure 12.1(a)) may be a complicated collection of particles. We can ask, however, about the probability that an incoming electron of energy E emerges with energy E', having been scattered through an angle θ, regardless of the state of these other particles. This probability is expressed by the differential cross-section $\mathrm{d}\sigma/\mathrm{d}\Omega\mathrm{d}E'$, where $\mathrm{d}E'$ represents a small range for the electron's final energy and $\mathrm{d}\Omega$ represents a small element of solid angle containing the direction of the outgoing electron (see appendix D). If the electron's kinetic energy is large enough, we can take its mass to be negligible, so $k^2 = k'^2 = m_\mathrm{e}^2 \approx 0$. The initial 4-momentum of the proton is $P^\mu = (M, \mathbf{0})$, where M is the proton mass. It is conventional to represent the energy lost by the electron, $E - E'$, and the scattering angle in terms of two Lorentz invariant quantities

$$\nu = M^{-1}(k - k') \cdot P = E - E' \tag{12.46}$$

$$q^2 = (k - k')^2 = -2k \cdot k' = -2EE' - 2|\boldsymbol{k}||\boldsymbol{k}'|\cos\theta = -4EE'\sin^2(\theta/2). \tag{12.47}$$

In fact, q is the 4-momentum carried by the virtual photon. A third variable

$$x = -q^2/2M\nu \tag{12.48}$$

will soon turn out to be useful. It has values in the range $0 < x < 1$, as can be shown by looking at the quantity

$$q^2(1 - x^{-1}) = q^2 + 2M\nu = q^2 + 2q \cdot P = (q + P)^2 - P^2 = P'^2 - P^2. \tag{12.49}$$

In the last expression, $P^2 = P_\mu P^\mu = M^2$ is the squared mass of the proton. The quantity $W = (P'^2)^{1/2} = (P'_\mu P'^\mu)^{1/2}$ is called the 'invariant mass' of the hadronic debris: it is the energy of this matter as measured in a frame where its net 3-momentum vanishes. It cannot be smaller than M, for if it were, the proton could spontaneously decay into this collection of particles, which we know does not happen. Since q^2 is negative, this implies that $x < 1$.

By using Lorentz invariance and the fact that the electromagnetic currents that interact with the photon are conserved, it is possible to show that the differential cross-section has the form

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega\mathrm{d}E'} = \frac{\alpha^2}{4E^2\sin^4(\theta/2)}\left[2W_1(q^2, \nu)\sin^2(\theta/2) + W_2(q^2, \nu)\cos^2(\theta/2)\right]. \tag{12.50}$$

The *structure factors* W_1 and W_2 depend on the internal structure of the proton and in general cannot be calculated reliably. Suppose, however, that the virtual photon interacts only with some point-like constituent inside the proton, as shown in figure 12.1(b), and that as far as the photon is concerned, this point-like particle can be considered in isolation from the rest of the proton as a free particle, say of mass m_p. This would have important implications, as we can see by considering energy–momentum conservation. Before colliding with the photon, the point particle is at rest (or very nearly so if its orbital motion inside the stationary proton is negligible), so its 4-momentum is $p^{\mu} = (m_{\mathrm{p}}, \mathbf{0})$. After the collision, its 4-momentum $p' = p + q$ satisfies $p'^2 = m_{\mathrm{p}}^2$, so we calculate

$$q^2 = (p' - p)^2 = p^2 + p'^2 - 2p \cdot p' = 2m_{\mathrm{p}}^2 - 2m_{\mathrm{p}} p'^0. \qquad (12.51)$$

Energy conservation also tells us that $\nu = E - E' = p'^0 - m_{\mathrm{p}}$, so we discover that $\nu = -q^2/2m_{\mathrm{p}}$. The structure functions must therefore be proportional to a δ function that enforces this constraint. In fact, if we assume that the point-like constituent is a spin-$\frac{1}{2}$ particle with charge Q (measured in units of e), then the structure functions can be worked out explicitly to be

$$W_1^{\text{point}}(q^2, \nu) = Q^2 \frac{-q^2}{4m_{\mathrm{p}}^2}\, \delta\left(\nu + \frac{q^2}{2m_{\mathrm{p}}}\right) \qquad (12.52)$$

$$W_2^{\text{point}}(q^2, \nu) = Q^2\, \delta\left(\nu + \frac{q^2}{2m_{\mathrm{p}}}\right). \qquad (12.53)$$

The factors of Q^2 take into account that in (12.50) there is one factor of α arising from the electron–photon vertex and another from the photon-proton vertex, which was assumed to refer to a particle of charge $Q = 1$.

If the virtual photon really did collide with a free, point-like particle inside the proton, the differential cross-section (12.50) would have a sharp spike at the particular value of the scattering angle θ consistent with the initial and final electron energies E and E'. This, however, is not what experiments find. The actual experimental situation can be represented reasonably well in terms of the *parton model*, which considers the point-like constituent (or parton) hit by the photon to carry some fraction of the proton's total 4-momentum, with a probability $f(\xi)\mathrm{d}\xi$ that this fraction is between ξ and $\xi + \mathrm{d}\xi$. In the case of an initially stationary proton, this means that the mass of the parton is $m_{\mathrm{p}} = \xi M$. Supposing that there are several species of partons, with charges Q_i and probability functions $f_i(\xi)$, we can calculate the structure function W_1 for the proton as

$$\begin{aligned} W_1(q^2, \nu) &= \sum_i Q_i^2 \int_0^1 \mathrm{d}\xi\; f_i(\xi)\; W_1^{\text{point}}(q^2, \nu)\Big|_{m_{\mathrm{p}} = \xi M} \\ &= \sum_i Q_i^2 \int_0^1 \mathrm{d}\xi\; \frac{-q^2}{4\xi^2 M^2} \delta\left(\nu + \frac{q^2}{2\xi M}\right) f_i(\xi) \\ &= \sum_i Q_i^2 \int_0^1 \mathrm{d}\xi\; \frac{-q^2}{4\xi M^2 \nu} \delta\left(\xi + \frac{q^2}{2M\nu}\right) f_i(\xi) \\ &= (2M)^{-1} \sum_i Q_i^2 f_i(x) \end{aligned} \qquad (12.54)$$

where, as defined above, $x = -q^2/2M\nu$. In the same way, we find

$$W_2(q^2, \nu) = \nu^{-1} \sum_i Q_i^2 x f_i(x). \tag{12.55}$$

The key result here is that the functions W_1 and νW_2 depend only on the single variable x rather than on q^2 and ν separately. This feature is known as *Bjorken scaling*, and it is brought about by the extra energy–momentum conservation constraint that comes into play when the virtual photon scatters elastically from a point-like constituent.

The parton model is not to be taken seriously as a theory of the internal structure of the proton; in particular, assigning a variable mass ξM to a fundamental particle makes little sense. Rather, it provides a rough-and-ready way of taking into account the interaction of a quark with the rest of the proton. (It can, however, be argued that the picture makes more sense when viewed from a frame of reference in which the proton has a very large energy and momentum, so that masses can be neglected and the parton simply carries a fraction ξ of the proton's energy.) With this reservation in mind, we might expect Bjorken scaling to become apparent in experimental data when the wavelength of the virtual photon is small enough for the internal structure of the proton to be resolved; that is, when $|q^2|$ is sufficiently large. In practice, the structure functions determined for fixed values of x are indeed found to be substantially independent of $|q^2|$ when $|q^2|$ is greater than about 1 GeV^2. Regardless of how literally we take the parton picture, this scaling provides clear evidence of the existence of quarks inside the nucleons. The scaling form of structure functions (12.54) and (12.55) together with similar functions for other scattering processes provide a framework for interpreting experimental data from which information about the quark content of nucleons can be extracted.

12.4.2 Quarks in the electroweak theory

As is apparent from the table at the beginning of this section, the quarks appear in pairs, (u, d), (c, s) and (t, b), whose charges differ by one unit. Like the (neutrino, charged-lepton) pairs, these are taken to form weak-isospin doublets. There is, however, a complication. The three gauge fields $\boldsymbol{W}_\mu$ form a weak-isospin triplet (see the discussion following (8.31)) but, as we have seen, W_μ^3 cannot be directly identified as the field operator for a particle because the term in (12.23) that generates the gauge boson masses involves a linear combination of W_μ^3 and B_μ. Now, the quark masses will be generated by a term in the Lagrangian density similar to (12.32), and the fields that appear in this term may, in general, be linear combinations of those needed to form the weak-isospin doublets. What these linear combinations are is a matter to be determined experimentally, and I shall shortly give a brief discussion of what is involved. The fact that no difficulty was encountered for leptons can be traced to the fact that all the neutrinos were assumed to have the same mass, namely zero. As with the leptons, then, the left-handed components of the various quarks are assembled into weak-isospin doublets, with $t = \frac{1}{2}$ and $y = \frac{1}{3}$, to give the correct charges:

$$\begin{pmatrix} u \\ d' \end{pmatrix}_{\mathrm{L}} \quad \begin{pmatrix} c \\ s' \end{pmatrix}_{\mathrm{L}} \quad \begin{pmatrix} t \\ b' \end{pmatrix}_{\mathrm{L}} \tag{12.56}$$

where d', s' and b' are linear combinations of d, s and b. All the right-handed components are SU(2) singlets, with hypercharge $y = \frac{4}{3}$ for u_{R}, c_{R} and t_{R} and $y = -\frac{2}{3}$ for d_{R}, s_{R} and b_{R}. The unprimed fields are those containing the creation and annihilation operators for particles of definite mass.

To see some of the implications of all this, let us construct the hadronic contribution to the charged current. It is helpful to express d', s' and b' in terms of d, s and b as

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{12.57}$$

where V is a unitary matrix called, after its inventors, the *Cabibbo–Kobayashi–Maskawa* (or CKM) *matrix*. This matrix can be written in terms of four *weak mixing angles*, analogous to the Weinberg angle. Following the pattern of (12.9), the hadronic charged current is

$$\mathcal{J}_{\mathrm{h}}^{\mu} = \bar{u}_{\mathrm{L}}\gamma^{\mu}d'_{\mathrm{L}} + \bar{c}_{\mathrm{L}}\gamma^{\mu}s'_{\mathrm{L}} + \bar{t}_{\mathrm{L}}\gamma^{\mu}b'_{\mathrm{L}} = (\bar{u}_{\mathrm{L}}, \bar{c}_{\mathrm{L}}, \bar{t}_{\mathrm{L}})\gamma^{\mu}V \begin{pmatrix} d_{\mathrm{L}} \\ s_{\mathrm{L}} \\ b_{\mathrm{L}} \end{pmatrix}. \tag{12.58}$$

The second form indicates that, had we also considered linear combinations of u, c and t, this would simply have meant redefining the matrix V. (More detailed arguments are necessary to show that V is unitary, however.)

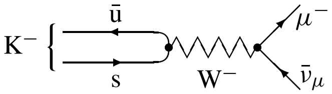
The situation is simpler if we ignore altogether the existence of the b quark (which was, indeed, unknown until about 1977) and the more recently discovered t quark. In that case, V is a 2×2 matrix, which can be parameterized by a single angle, the *Cabibbo angle* θ_{C}:

$$V = \begin{pmatrix} \cos\theta_{\mathrm{C}} & \sin\theta_{\mathrm{C}} \\ -\sin\theta_{\mathrm{C}} & \cos\theta_{\mathrm{C}} \end{pmatrix}. \tag{12.59}$$

The hadronic charged current becomes

$$\mathcal{J}_{\mathrm{h}}^{\mu} = (\bar{u}_{\mathrm{L}}\gamma^{\mu}d_{\mathrm{L}} + \bar{c}_{\mathrm{L}}\gamma^{\mu}s_{\mathrm{L}})\cos\theta_{\mathrm{C}} + (\bar{u}_{\mathrm{L}}\gamma^{\mu}s_{\mathrm{L}} - \bar{c}_{\mathrm{L}}\gamma^{\mu}d_{\mathrm{L}})\sin\theta_{\mathrm{C}}. \tag{12.60}$$

Consider, for example, the decay of a K$^-$ meson, whose quark content is (ūs), into a negative muon and an antineutrino (K$^- \to \mu^- + \bar{\nu}_{\mu}$). What happens, according to the GWS theory, is that the quark and antiquark annihilate to produce a virtual W$^-$, which subsequently decays to produce the leptons:



The field W_μ^+, which creates the W$^-$, couples to the hadronic current (12.60), in which the operator $\bar{u}_L\gamma^\mu s_L$ that annihilates the quarks has the coefficient $\sin\theta_C$. Thus, the ūsW$^-$ vertex has a factor of $\sin\theta_C$ and the decay rate a factor of $\sin^2\theta_C$. If there were no mixing, or, in other words, if d' and s' were identical with d and s, the decay could not take place. In terms of the Fermi theory, the K$^-$ decay can be thought of as involving an effective Fermi constant $G_F\sin\theta_C$. Unfortunately, the actual value of the decay rate depends on details of the strong interaction mechanism that binds the ū and s quarks to form the K$^-$, so we cannot use it directly to determine θ_C. An estimate of θ_C can be made if we assume, for example, that this mechanism gives the K$^-$ and π^- the same structure, apart from the fact that π^- is made from ū and d. In that case, the matrix elements T_{fi} (see appendix D) for the decays K$^- \to \mu^- + \bar{\nu}_\mu$ and $\pi^- \to \mu^- + \bar{\nu}_\mu$ should have the ratio $\tan\theta_C$. Taking account also of the kinematical factors that influence the decay rates, one finds that $\sin\theta_C \approx 0.22$. In general, although the CKM matrix is a property of the weak interactions, we see that its elements can be deduced from experimental data only if information about strong-interaction matrix elements is available. Mostly, this information can be obtained only by means of special assumptions or simplified models.

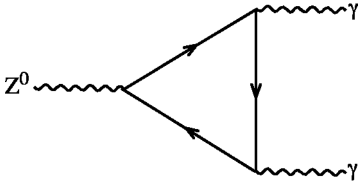


Figure 12.2. Example of a diagram which causes anomalies. Diagrams which contain this as a subdiagram cannot, in general, have their infinities removed by renormalization.

12.5 Colour and Quantum Chromodynamics

Although the GWS model as I have described it so far has a gauge-invariant action, it is not renormalizable. This is because of the occurrence of *anomalies*, which were mentioned in chapter 9. An example of part of a Feynman diagram whose divergence cannot be renormalized away is shown in figure 12.2. The theory will be renormalizable if the net contribution of all diagrams of this type is zero. Now, one such diagram can be formed with each charged fermion species circulating in the closed loop and, as it turns out, the condition for the divergences to cancel is that the sum of the charges of all these species is zero. In the standard model, this is true if two conditions are met. The first is that the fermion species fall into a number of complete *families* or *generations*, each family comprising a neutrino, a negatively-charged lepton and a pair of quarks with charges of $\frac{2}{3}$ and $-\frac{1}{3}$. Evidently, the known fermions do fall into just three such families, namely (ν_{e}, e, u, d), (ν_μ, μ, c, s) and (ν_τ, τ, t, b). The second condition is that each quark flavour should count as *three* species. In fact, it is believed that each flavour does indeed correspond to three distinct species, all having the same mass and electroweak properties, but distinguished by a property called *colour*. There is no universal agreement on the three colours used to label these species, but the primary colours red, green and blue are commonly used.

The earliest reason for this hypothesis was that some baryons appeared to consist of three identical quarks in a symmetric state, which is at variance with the fermionic nature of the quarks. This no longer presents a problem if the three quarks, while having the same flavour, are of different colours. Direct evidence for the existence of three colours comes from several sources. The neutral pion π^0 is an antisymmetric combination of $\mathrm{u\bar{u}}$ and $\mathrm{d\bar{d}}$ bound states, which decays to two photons via a Feynman diagram similar to figure 12.2. In this case, the integral turns out to be finite, but it is proportional to the number of quark species circulating in the loop, and gives the correct answer for the lifetime of the π^0 only when allowance is made for three colours. In high-energy collisions of electrons and positrons, these two particles annihilate to form a virtual photon, which may subsequently decay into particle-antiparticle pairs of any fermion species that can be created with the energy available. One possibility is that these particles are muons ($\mathrm{e}^+e^- \to \mu^+\mu^-$), which can be detected directly. Another possibility is the formation of quark-antiquark ($\mathrm{q\bar{q}}$) pairs which, as with deep inelastic scattering, are eventually manifested as a complicated collection of hadrons. The total probability for the formation of $\mathrm{q\bar{q}}$ pairs is proportional to $\sum_i Q_i^2$, where the sum is over all quark species that can be produced at a given energy. If each flavour of quark comes in N_{c} colours, all with the same electric charge, then this becomes $N_{\mathrm{c}}\sum_f Q_f^2$, where the sum is over quark flavours. According to the parton model, the ratio of the probabilities for forming hadrons or muons, as measured by the corresponding cross-sections, is just

$$\frac{\sigma(\mathrm{e}^+\mathrm{e}^- \to \mathrm{hadrons})}{\sigma(\mathrm{e}^+\mathrm{e}^- \to \mu^+\mu^-)} = N_{\mathrm{c}}\sum_f Q_f^2 \qquad (12.61)$$

the muons having $Q = 1$. Apart from details that are not accounted for by the parton model, this agrees well with experimental data, provided that we take $N_{\mathrm{c}} = 3$.

The existence of three quark colours provides the basis of the current theory of strong interactions, known as *quantum chromodynamics* or QCD. Here, I can do no more than outline some of its essential features. The three colours of a given quark flavour are taken to form a basic triplet

$$u = \begin{pmatrix} u_{\mathrm{r}} \\ u_{\mathrm{g}} \\ u_{\mathrm{b}} \end{pmatrix} \qquad d = \begin{pmatrix} d_{\mathrm{r}} \\ d_{\mathrm{g}} \\ d_{\mathrm{b}} \end{pmatrix} \quad \text{etc.} \tag{12.62}$$

The set of unitary transformations $u \to \exp[\frac{1}{2}\mathrm{i}\boldsymbol{\alpha}(x) \cdot \boldsymbol{\lambda}]u$, which rearrange the three colours amongst themselves, constitutes the *colour gauge group* SU(3). This group has eight generators. That is, there are eight linearly independent, Hermitian λ matrices, analogous to the Pauli matrices of SU(2). Consequently, when this group is used to construct a gauge theory, there are eight independent gauge fields and eight associated gauge bosons. These are called *gluons*, being held to form the 'glue' that binds quarks into hadrons. Like the quarks, gluons are (it seems) permanently confined inside the hadrons. Direct evidence for their existence can be gleaned from the structure functions of deep inelastic scattering. The functions $f_i(x)$ in (12.54) and (12.55) represent the probabilities that the ith constituent species carries a fraction x of the proton's total momentum. The total fraction carried by all the constituents must obviously be 1, which implies that $\sum_i \int_0^1 \mathrm{d}x\, x f_i(x) = 1$. However, when the $f_i(x)$ deduced from measured structure functions are inserted into this 'sum rule', a shortfall of about 50% is found. The implication is that some 50% of the momentum is carried by electrically neutral constituents, which do not interact with the virtual photon. If QCD is correct, then these neutral constituents can be identified as gluons.

Unlike the electroweak theory, QCD contains no Higgs fields, so the gluons are massless. It might therefore appear that the colour forces should, like electromagnetic forces, have a long range and be easily detectable in the laboratory. It is believed, however, that QCD possesses a property known as *confinement*. The potential energy of two quarks increases linearly with the distance between them. Thus, if we try to separate, say, the quark and antiquark in a pion, the increase in potential energy eventually favours the formation of a new quark-antiquark pair and we obtain not two widely separated quarks but two widely separated mesons. Only bound states which have no net colour (colour singlets) have a finite energy and this, in outline, explains why isolated quarks and gluons are never observed. The very different properties of QCD and QED can be traced to the non-Abelian nature of SU(3). As we saw in chapter 8, this implies that the gluons themselves carry a colour 'charge' and thus interact directly with each other, in contrast to photons, which are electrically neutral.

While few theorists doubt the validity of this picture, it has not, as far as I know, been possible to give a definitive proof. The difficulty is that perturbation theory cannot be used. Perturbation theory, after all, assumes that the field operators in the theory can, to a first approximation, be interpreted as creation and annihilation operators for observable, free particles, and in QCD this is not true. It has proved fruitful to consider an approximate theory in which spacetime is replaced by a discrete four-dimensional lattice of points, quite analogous to the lattice models of statistical mechanics. For such *lattice gauge theories*, the confinement property can be proved, but the proof does not necessarily remain valid when the lattice spacing is taken to zero. If spacetime is approximated not only as a discrete set of points, but also as being of finite extent, then functional integrals such as we encountered in earlier chapters reduce to ordinary multiple integrals, whose values can be estimated numerically. This idea provides an alternative means of approximation when perturbation theory is inapplicable; in fact, it is the only known practical method of estimating quantities such as the mass of a proton directly from QCD. This method of approximation has its own difficulties. One is that, although the lattice has only a finite number of points, this number must be very large if the lattice is to provide a reasonable approximation to a spacetime continuum, and to represent a region of spacetime large enough to contain several hadrons. The computing power needed to deal with lattices of sufficient size is, even by present-day standards, enormous. Another is that representing fermions correctly in the lattice approximation turns out to be quite tricky. Nevertheless, at the time of writing, it has become possible to estimate the masses of the lighter hadrons (specifically, those containing u, d and s quarks in what amounts to their 'ground state') with an accuracy that reproduces experimental data to within 10% or better. It is also possible to estimate from first principles the strong-interaction matrix elements that are needed to extract information on the CKM matrix from measured decay rates and scattering cross-sections, although such calculations are less well advanced.

The confinement of quarks (or, more accurately, of colour) is a large-distance or low-energy phenomenon. At high energies, QCD has the complementary property of *asymptotic freedom*. This means that the running coupling constant $\alpha_{\mathrm{s}}(-q^2)$, the strong-interaction equivalent of the energy-dependent fine structure constant (9.92), becomes very small at high energies. In fact, the result analogous to (9.94) for its high-energy behaviour is

$$\alpha_{\mathrm{s}}(-q^2) = \alpha_{\mathrm{s}}(\mu^2)\left[1 + (11 - \tfrac{2}{3}n_{\mathrm{f}})\frac{\alpha_{\mathrm{s}}(\mu^2)}{4\pi}\ln\left(\frac{-q^2}{\mu^2}\right)\right]^{-1} \tag{12.63}$$

where n_{f} is the number of quark flavours. How this behaves for large values of $-q^2$ clearly depends on the sign of the quantity $(11 - \frac{2}{3}n_{\mathrm{f}})$. The contribution $-\frac{2}{3}n_{\mathrm{f}}$ arises from the effect of quark-antiquark pairs in screening the strong 'charge' of a particle, and is entirely analogous to the vacuum polarization in QED that we discussed in §9.7.4. The positive term, 11, comes from the self-interaction of gluons, which has no analogue in QED. It results from the non-Abelian nature of the SU(3) colour gauge group, which, as in the SU(2) theory we studied in chapter 8, leads to the presence of nonlinear terms in the field strength (8.37). Provided that there are no more than 16 quark flavours (and only 6 are known), this self-interaction of gluons is the more important effect, and we see that it causes $\alpha_{\mathrm{s}}(-q^2)$ to decrease with increasing values of $-q^2$. Conversely, $\alpha_{\mathrm{s}}(-q^2)$ becomes very large at low energies—a fact which might seem intuitively consistent with confinement, but is not in fact sufficient to show that confinement

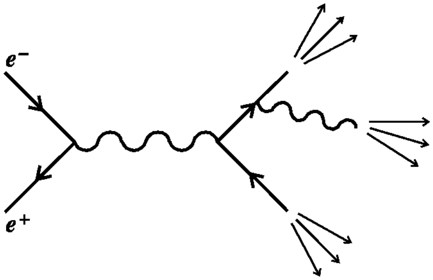


Figure 12.3. Schematic view of a 3-jet event produced in an electron-positron collision. Roughly collimated jets of particles emerge in the directions of a quark, an antiquark and a gluon formed in the initial decay of a virtual photon.

actually occurs. Amongst other things, this means that there is no QCD equivalent of 'the' fine structure constant α, which measures the electronic charge apparent at macroscopic distances and is the low-energy limit of $\alpha(-q^2)$. Because of this, it has become conventional to parameterize the strength of colour forces by an energy scale Λ_{QCD}. At the level of approximation I am using here, we can write

$$\alpha_{\mathrm{s}}(\mu^2) = 4\pi \left[(11 - \tfrac{2}{3}n_{\mathrm{f}}) \ln(\mu^2/\Lambda_{\mathrm{QCD}}^2) \right]^{-1} \tag{12.64}$$

and the energy-dependent coupling constant becomes

$$\alpha_{\mathrm{s}}(-q^2) = 4\pi \left[(11 - \tfrac{2}{3}n_{\mathrm{f}}) \ln(-q^2/\Lambda_{\mathrm{QCD}}^2) \right]^{-1} . \tag{12.65}$$

All reference to the renormalized coupling $\alpha_{\mathrm{s}}(\mu^2)$, defined at arbitrary, but fixed energy scale μ has disappeared and the intrinsic strength of the interactions is characterized instead by Λ_{QCD}. The fact that a dimensionless coupling can be replaced with a parameter having the dimensions of energy is sometimes referred to as *dimensional transmutation*.

Because $\alpha_{\mathrm{s}}(-q^2)$ is small at high energies, perturbation theory can be applied to good effect in understanding processes such as deep inelastic scattering. By comparing calculated structure functions with those measured experimentally, it has been possible, for example, to confirm the energy dependence of α_{s} and to account for departures from Bjorken scaling that are observed at small values of x.

A striking feature of high-energy data is the formation of *jets* of hadronic particles. These are interpreted as signalling the ejection from a nucleon of individual quarks or gluons, which subsequently acquire, through the creation of particle-antiparticle pairs, the partners needed to form a shower of colourless hadrons. The total momentum of particles in the jet is the momentum that originally belonged to a single quark or gluon (see figure 12.3). By observing

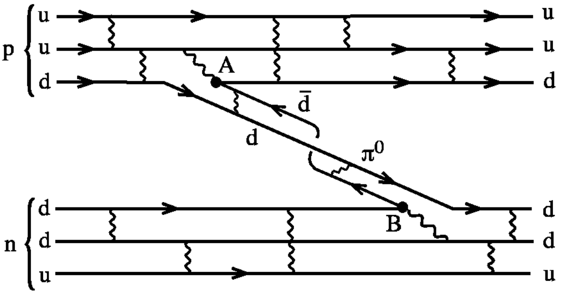


Figure 12.4. Contribution to the force between a proton and a neutron due to exchange of a π^0. Quarks are bound into hadrons by the exchange of gluons. At A, a gluon decays to form a d$\bar{\text{d}}$ pair and at B a d$\bar{\text{d}}$ pair annihilates to form a gluon. The net effect is the exchange of a π^0. Backward-pointing arrows denote a forward-moving antiquark.

the production of jets in, for example, e^+e^- and proton-antiproton collisions, it is possible, in effect to study the scattering of individual quarks and gluons and perturbative QCD accounts for much of this data with impressive accuracy.

It should be emphasized that QCD describes the strong interactions that bind quarks inside the observed hadrons. The forces that act between these hadrons, for example, those which bind protons and neutrons to form atomic nuclei or account for the low-energy scattering of protons and neutrons, should also have their origins in QCD, but they cannot be attributed to exchange of gluons. Figure 12.4 illustrates, in terms of the flow of quarks, how the force between a proton and neutron can be attributed to the exchange of, for example, a neutral pion. The fundamental origin of the force is the QCD interaction, which binds quarks in all three hadrons and causes the creation and annihilation of quark-antiquark pairs. However, their net effect at low energies or large distances can be modelled by treating the pion as a fundamental spin-0 particle. This leads to a *one-particle exchange potential*, which has the Yukawa form (9.85). As I indicated in chapter 9, the pion mass corresponds to a range for this effective force that is characteristic of the separation of nucleons in a nucleus or, indeed, of the size of a nucleon. This simple model has rather restricted applicability, though. To improve on it, account must be taken of other mesons that might be exchanged and of the internal structure of these particles.

12.6 Grand Unified Theories

The gauge theory whose construction I have outlined so far constitutes the *standard model* of particle physics. Within the uncertainties involved in actually calculating quantities that can be directly compared with experimental data, it appears to be consistent with all known phenomena (except, perhaps, for the possibility that neutrinos may, after all, have small masses). From a theoretical point of view, it is nevertheless held to be unsatisfactory, partly because it contains a large number of parameters which simply have to be adjusted to values determined by experiment, and partly because it does not represent a truly unified description of the fundamental forces. I shall give just two examples of the improvements that might be sought.

The first concerns the question of *charge quantization*. We saw in chapter 8 that the numbers λ_i (in (8.17), for example), which express the charges of different particles as multiples of the fundamental charge e could have any values. There is no explanation for the fact that they are observed to have integer or, in the case of quarks, simple rational values. In the GWS electroweak theory, the charges of particles belonging to an SU(2) doublet must differ by one unit, but the hypercharge of each multiplet, which gives the actual charges through the Gell-Mann–Nishijima formula, is assigned simply to fit the observed facts.

The second unsatisfactory feature is that the standard model involves three independent gauge coupling constants, namely the g and g' of the electroweak theory and a third, g_s, for QCD. This is because the gauge symmetry group is SU(3)×SU(2)×U(1), which means that the SU(3) transformations that rearrange colours, the SU(2) weak-isospin rotations and the U(1) phase transformations all act independently of each other. It is, of course, satisfying that the strong, weak and electromagnetic interactions, which at first sight have very different physical effects, can all be described in essentially the same terms as gauge theories. Moreover, the weak and electromagnetic interactions are intimately related in the GWS theory. Indeed, the relative weakness of the weak interactions, as measured by the Fermi constant G_F, is seen from (12.38) to be due to the relatively large masses of the gauge bosons rather than to the size of the coupling constant g, which is actually greater than e. This and the different ranges of the two interactions are seen to be consequences of spontaneous symmetry breaking. That having been said, we still need three coupling constants to account for the three forces. In the view of most theorists, it would be much more satisfactory if we could account for all three forces using a single coupling constant, with all the differences arising from spontaneous symmetry breaking. In particular, we would like to be able to *predict* the value of the Weinberg angle which, according to (12.27), just measures the ratio of g and g'.

Considerations such as these have led to the invention of *grand unified theories*, whose principal feature is that the fundamental gauge group should be *simple*. This means that it cannot be expressed as the product of several independent groups, which immediately implies the existence of only a single gauge coupling constant. The earliest and simplest of these theories was invented by H Georgi and S Glashow (1974), who took the gauge group to be SU(5). The 15 fermions of a single family (counting colours and left- and right-handed components separately for this purpose) fit into two SU(5) multiplets, of which the simpler is

$$\begin{pmatrix} \nu_{\mathrm{eL}} \\ e_{\mathrm{L}} \\ d^{\mathrm{c}}_{\mathrm{rR}} \\ d^{\mathrm{c}}_{\mathrm{gR}} \\ d^{\mathrm{c}}_{\mathrm{bR}} \end{pmatrix}. \tag{12.66}$$

In this notation, $d^{\mathrm{c}}_{\mathrm{rR}}$, for example, denotes the charge conjugate of the right-handed component of the field operator for a red down quark. The charge conjugate of a right-handed component is left handed (see exercise 7.8), so all the field operators are in fact left handed. In terms of particles, the electron and its neutrino are grouped with the anti-down quark, whose charge is $+\frac{1}{3}$.

The gauge transformations that act on this multiplet are of the form $\exp[\frac{1}{2}\mathrm{i}\boldsymbol{\alpha}(x)\cdot\boldsymbol{\xi}]$, where the matrices ξ^a are the SU(5) analogues of the Pauli matrices. There are 24 of these matrices, which are Hermitian 5×5 matrices whose trace is zero. The standard model is included in the SU(5) model, because some of these transformations correspond to those of SU(3)×SU(2)×U(1). For example, three of the ξ^a can be written as

$$\begin{pmatrix} \multicolumn{2}{c}{\tau^a} & 0 & 0 & 0 \\ & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{12.67}$$

where τ^a are the Pauli matrices. These generate the weak-isospin transformations of the electron–neutrino doublet, leaving the right-handed quarks unchanged. A further eight are

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & & & \\ 0 & 0 & & \lambda^a & \\ 0 & 0 & & & \end{pmatrix} \tag{12.68}$$

λ^a being the SU(3) matrices, which generate colour transformations of the quarks without affecting the leptons. We must think a little more carefully about the U(1) phase transformations, however. In any grand unified theory, all the particles belonging to one standard-model family must fill up a complete multiplet of the grand-unified gauge group (although this may be reducible into sub-multiplets, as in SU(5)). Let us assemble all 15 fields into a column matrix ψ. In the language of the standard model, its gauge-covariant derivative is a generalization of the one that appears in (12.16), namely

$$\mathrm{D}_\mu\psi = \left(\partial_\mu + \mathrm{i}gW^3_\mu T^3 + \mathrm{i}g'B_\mu\tfrac{1}{2}Y + \ldots\right)\psi \tag{12.69}$$

where T^3 is a diagonal 15×15 matrix whose elements are the t^3 components of weak isospin for the various particles, Y is another diagonal 15×15 matrix whose elements are the hypercharges, and ... are the remaining generators of the grand-unified gauge group together with their associated gauge fields. In grand-unified language, it must be possible to write this as

$$\mathrm{D}_\mu\psi = \left(\partial_\mu + \mathrm{i}g_2 W_\mu^3 T^3 + \mathrm{i}g_1 B_\mu T^0 + \ldots\right)\psi \tag{12.70}$$

where both T^3 and T^0 are generators of the grand-unified gauge group. The coupling constant $g_2 \equiv g$ belongs to the SU(2) sector of the standard model and g_1 belongs to the U(1) sector. Eventually, we shall have to set $g_1 = g_2 = g_\mathrm{G}$, where g_G is the single coupling constant of the grand-unified theory, but we shall see that there are good reasons for keeping them separate at this stage. To make these two expressions equivalent, let us say that $g' = c^{-1}g_1$ and $Y = 2cT^0$, where c is a constant. The value of this constant is determined by the fact that the two generators T^3 and T^0 must satisfy the normalization condition (8.40). In particular, we must have $\mathrm{Tr}(T^3)^2 = \mathrm{Tr}(T^0)^2 = (1/4c^2)\,\mathrm{Tr}\,Y^2$, or

$$\sum_i (t_i^3)^2 = \frac{1}{4c^2}\sum_i y_i^2 \tag{12.71}$$

where t_i^3 and y_i are the weak-isospin components and hypercharges of the particles in the multiplet. We have to take account of: ν_{eL} and e_L with $t^3 = \pm\frac{1}{2}$ and $y = -1$; e_R with $t^3 = 0$ and $y = -2$; u_L and d_L with $t^3 = \pm\frac{1}{2}$ and $y = \frac{1}{3}$; u_R with $t^3 = 0$ and $y = \frac{4}{3}$; d_R with $t^3 = 0$ and $y = -\frac{2}{3}$. Each quark counts as 3 species, on account of its colour. The fact that the right-handed particles are represented by their left-handed antiparticles as in (12.66) doesn't matter for this purpose, because their quantum numbers are squared. Using these values in (12.71), we find that $c = (5/3)^{1/2}$.

We can now derive two simple consequences of grand unification. First, the electromagnetic field A_μ must be a gauge field of the grand-unified theory. It appears in a covariant derivative such as (12.69) or (12.70) in the combination $\mathrm{i}eA_\mu Q$, where the diagonal matrix $Q = T^3 + \frac{1}{2}Y = T^3 + cT^0$ has elements which are the charges of all the particles in whichever multiplet we choose to look at, measured in units of e. In the SU(5) theory, all the generators are Hermitian matrices whose trace is zero, so the charges of all the particles in any multiplet must add to zero. Applying this principle to the multiplet (12.66), we see that the charge of an anti-d quark must be exactly $-\frac{1}{3}$ of the charge of the electron. In the case of the second multiplet, which contains the u quark, a similar argument shows that u has a charge which is exactly $-\frac{2}{3}$ of the electronic charge. Thus, the SU(5) theory provides an explanation of the fact that the charge of the proton is exactly -1 times the charge of the electron; this *charge quantization* is a major success of grand unification.

The second consequence comes about when we set $g_1 = g_2 = g_\mathrm{G}$ or $g_2 = g_\mathrm{G}$ and $g' = c^{-1}g_\mathrm{G}$. In view of (12.27), this gives us a *prediction* for the weak mixing angle

$$\sin^2\theta_{\mathrm{W}} = \frac{g'^2}{g^2 + g'^2} = \frac{1}{1+c^2} = \frac{3}{8} = 0.375. \tag{12.72}$$

Compared with the experimental value (12.43), this does not seem like an unqualified success, but we have yet to take account of two important ingredients, *viz.* spontaneous symmetry breaking and the running of coupling constants with energy.

The SU(5) theory has 24 symmetry generators and therefore 24 gauge bosons; other grand unified theories (or GUTs) may have more. Of these, 12 can be identified with the gauge bosons of the standard model, but the rest, which I shall denote collectively by X, are unknown to experimenters. If a GUT has indeed been used by nature, then these extra gauge bosons must be very heavy, or else their existence would upset the success of the standard model. The GUT symmetry must, it seems, be broken at two levels, by two sets of Higgs fields. One stage of symmetry breaking gives a large mass, say M_{X}, to the X bosons while leaving the SU(3)×SU(2)×U(1) symmetry intact and the standard-model gauge bosons massless. This symmetry can then be spontaneously broken in the way we have already seen, leaving only the U(1) symmetry of electromagnetism. What does this imply for the running coupling constants? At energies greater than M_{X}, the effects of spontaneous symmetry breaking will not matter greatly and all the physics will be controlled by a single coupling constant $g_{\mathrm{G}}(Q^2)$. (Here, I will use the conventional notation $Q^2 = -q^2$, because there will be no danger of confusing this Q with an electric charge.) At energies well below M_{X}, propagators for the X bosons will be very small. It should be possible to ignore these particles for most purposes, and physics should be essentially the same as in the standard model, with its three coupling constants g_1, g_2 and g_3, the last of these being the QCD coupling. At energies close to M_{X}, these two descriptions must become equivalent. Thus, as illustrated in figure 12.5, we should have $g_1(M_{\mathrm{X}}^2) = g_2(M_{\mathrm{X}}^2) = g_3(M_{\mathrm{X}}^2) = g_{\mathrm{G}}(M_{\mathrm{X}}^2)$, although the same equalities need not hold at lower energies. The running of the three coupling constants at low energies is, according to this argument, governed just by the standard model, and is independent of any special assumptions about the nature of the hypothetical GUT. We can therefore use standard-model data to test whether they actually do become equal and, if so, at what energy.

Let us explore this question in a simple approximation. Defining $\alpha_i = g_i^2/4\pi$, the running coupling constants are found, at the first order of perturbation theory, to be given by

$$\alpha_i^{-1}(Q^2) = \alpha_i^{-1}(M_{\mathrm{Z}}^2) + \frac{\beta_i}{4\pi}\ln\left(\frac{Q^2}{M_{\mathrm{Z}}^2}\right) \tag{12.73}$$

the reference scale $\mu^2 = M_{\mathrm{Z}}^2$ being experimentally rather well defined. The

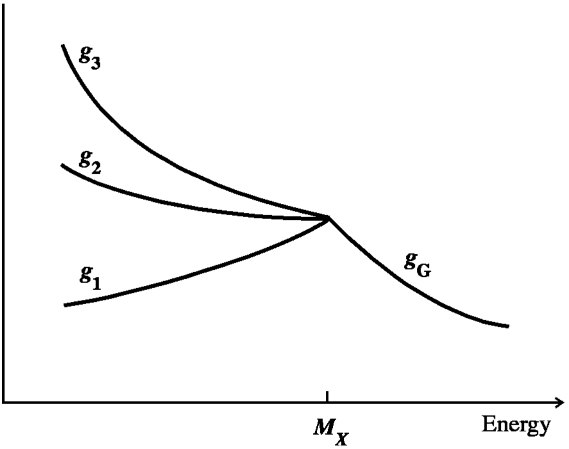


Figure 12.5. Energy dependence of the running coupling constants in a grand-unified theory. The spontaneously broken symmetry which gives the U(1), SU(2) and SU(3) couplings at low energies is restored at an energy approximately equal to the typical *X* boson mass.

constants β_i are

$$\beta_1 = -\tfrac{4}{3}n_g - \tfrac{1}{10}n_h = -\tfrac{41}{10} \qquad (12.74)$$

$$\beta_2 = \tfrac{22}{3} - \tfrac{4}{3}n_g - \tfrac{1}{6}n_h = \tfrac{19}{6} \qquad (12.75)$$

$$\beta_3 = 11 - \tfrac{4}{3}n_g = 7 \qquad (12.76)$$

where n_g is the number of families (or generations) of quarks and leptons and n_h is the number of Higgs doublets. I have taken $n_g = 3$ and $n_h = 1$. The standard-model data we have at our disposal are the QCD coupling strength α_3, the electromagnetic fine-structure constant α and the mixing angle θ_W, whose values at $Q^2 = M_Z^2$ I shall denote by a circumflex. These values are determined experimentally as

$$\hat{\alpha}_3 = 0.12 \qquad \hat{\alpha}^{-1} = 128.9 \qquad \sin^2\hat{\theta}_W = 0.232. \qquad (12.77)$$

We can make direct use of the fine-structure constant by using (12.31) to write it as $\alpha^{-1} = c^2\alpha_1^{-1} + \alpha_2^{-1}$, which implies that its energy dependence is given by

$$\alpha^{-1}(Q^2) = \hat{\alpha}^{-1} + \frac{\beta}{4\pi}\ln\left(\frac{Q^2}{M_Z^2}\right) \qquad (12.78)$$

with $\beta = c^2\beta_1 + \beta_2 = -11/3$. One way of phrasing our question is now the following. Using only the experimental values of $\hat{\alpha}$ and $\hat{\alpha}_3$, we can estimate the unification energy M_X, at which we expect $\alpha_1 = \alpha_2 = \alpha_3$, by solving the equation

$\alpha^{-1}(M_{\mathrm{X}}^2) = (1+c^2)\alpha_3^{-1}(M_{\mathrm{X}}^2)$. Then, taking the value (12.72) of $\sin^2\theta_{\mathrm{W}}$ to be the one that applies at $Q = M_{\mathrm{X}}$, we can use the running coupling constants to obtain a revised prediction for $\sin^2\hat{\theta}_{\mathrm{W}}$. If this agrees with the measured value, it would indicate that all three coupling constants are related in the way that grand unification requires.

The first step of this calculation gives

$$\ln\left(\frac{M_{\mathrm{X}}^2}{M_{\mathrm{Z}}^2}\right) = \frac{4\pi\left[\hat{\alpha}^{-1} - (1+c^2)\hat{\alpha}_3^{-1}\right]}{(1+c^2)\beta_3 - \beta} \tag{12.79}$$

or $M_{\mathrm{X}} \approx 1.1 \times 10^{13} M_{\mathrm{Z}} \approx 10^{15}$GeV. To obtain our prediction for $\sin^2\hat{\theta}_{\mathrm{W}}$, we use (12.27) and (12.31) to write $\sin^2\theta_{\mathrm{W}} = e^2/g^2$ and hence

$$\sin^2\theta_{\mathrm{W}}(Q^2) = \frac{\alpha(Q^2)}{\alpha_2(Q^2)} = \frac{\sin^2\hat{\theta}_{\mathrm{W}} + (\beta_2/4\pi)\hat{\alpha}\ln(Q^2/M_{\mathrm{Z}}^2)}{1 + (\beta/4\pi)\hat{\alpha}\ln(Q^2/M_{\mathrm{Z}}^2)}. \tag{12.80}$$

Setting $\sin^2\theta_{\mathrm{W}}(M_{\mathrm{X}}^2) = \frac{3}{8}$, we can solve this to get the prediction $\sin^2\hat{\theta}_{\mathrm{W}} \approx 2.07$. This is encouragingly close enough to the measured value (12.43), but certainly not within the experimental uncertainty. Of course, the calculation was only approximate; it could be improved by including contributions to the running coupling constants from higher orders of perturbation theory. However, a different route has been followed in practice, which is to calculate all three of the running coupling constants $\alpha_i(Q^2)$ using the standard-model data as initial conditions at $Q = M_{\mathrm{Z}}$. The most accurate calculations indicate that although any two of the α_i become equal at an energy close to 10^{15}GeV, they do not all become equal at exactly the same point. This may be an indication that, while the general idea of grand unification is plausible, some significant ingredient is missing.

By comparison with the W and Z masses of about 10^2 GeV, or with energies of the order of 10^3 GeV that can be produced by present-day accelerators, the unification energy of 10^{15} GeV is enormous. We have no hope of observing the X particles directly, and any indirect effects that their existence might bring about will be very small. One such effect, which could in principle be observed, is *proton decay*. In the standard model, the currents that couple to the weak gauge fields contain only terms of the form $\bar{q}\gamma^\mu q$ or $\bar{\ell}\gamma^\mu\ell$, where q and ℓ generically denote quarks and leptons. It follows that a quark can be transformed into a quark of a different flavour by emitting a weak gauge boson, but not into a lepton. Consequently, a baryon can decay only into a lighter baryon, together with a virtual weak boson, which subsequently produces a lepton-antilepton pair, as in the beta decay of a free neutron. The proton, being the lightest baryon, cannot decay at all. The reason for this is that quarks and leptons are contained in separate SU(2) multiplets. Each multiplet of a GUT, however, contains both quarks and leptons. Therefore, the currents that couple to X gauge fields contain terms of the form $\bar{q}\gamma^\mu\ell$ and $\bar{\ell}\gamma^\mu q$, which permit the transformation of a quark into a lepton by the emission of an X boson. Moreover, the GUT multiplets may

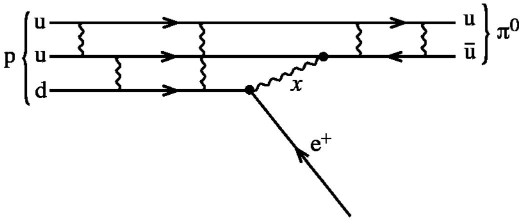


Figure 12.6. A contribution to the decay of a proton, producing a positron and a π^0. The X boson has a charge of $-4e/3$.

contain both left-handed components of quark fields and the charge conjugates of their right-handed components, and this permits the transformation of a quark into an antiquark. Because of this, proton decay becomes possible, and figure 12.6 shows one mechanism whereby it can decay into a π^0 and a positron. A simple estimate of the proton's lifetime can be made from the formulae of appendix D, together with dimensional analysis. The matrix element T_{fi} for the emission and absorption of an X boson is proportional to $\alpha_{\mathrm{G}} M_{\mathrm{X}}^{-2}$, as in our estimate (12.8) of the Fermi constant. The decay rate Γ has the dimensions of energy in natural units, and must be proportional to $(\alpha_{\mathrm{G}} M_{\mathrm{X}}^{-2})^2 M_{\mathrm{p}}^5$, because the proton mass M_{p} is the only relevant energy scale. Up to a numerical factor, we therefore estimate the proton's lifetime as

$$\tau_{\mathrm{p}} = \hbar\Gamma^{-1} \sim \hbar\alpha_{\mathrm{G}}^{-2} M_{\mathrm{X}}^4 M_{\mathrm{p}}^{-5} \sim 10^{38}\,\mathrm{s} \sim 10^{31}\,\text{years} \qquad (12.81)$$

where the factor of $\hbar$ converts units of energy^{-1} into seconds, and the grand-unified coupling is taken as $\alpha_{\mathrm{G}} \approx 0.1$. A more detailed calculation based on the SU(5) theory produces much the same result. Clearly, proton decays will be very rare. To put this lifetime in context, the current age of the universe is only some 10^{10} years. On the other hand, since the proton's mass is $M_{\mathrm{p}} \approx 1.67 \times 10^{-27}$ kg, we might hope to detect one or two decays per year by keeping some 10^4 kg of a suitable material under observation. Several experiments of this kind have been undertaken, usually deep underground to avoid the intrusion of cosmic radiation. No decays have been observed, and the experimental limit on the proton's lifetime is that τ_{p} is no smaller than about 10^{32} years. It is an odd fact that this experimental limit is of the same order of magnitude as the actual lifetime expected on the basis of grand unified theories. There is, of course, some uncertainty in the predicted lifetime, but experts are more or less agreed that these experiments rule out the SU(5) theory as a model of the real world. Many other GUTs can be devised, though, and some of them predict longer-lived protons.

Clearly, the value of grand unified theories lies much more in their aesthetic appeal in providing a completely unified description of the three interactions, and suggesting an explanation for charge quantization, than in their utility for interpreting hard experimental data. Even their aesthetic appeal has its limitations.

In the SU(5) theory, for example, the observed fermions have to be fitted into two multiplets, and it is hard to see any good physical reason for treating the particles in (12.66) on a different footing from the others. Similarly, in order to reproduce the successes of the standard model, we had to introduce two stages of spontaneous symmetry breaking, using two sets of Higgs fields. This is simply an *ad hoc* manoeuvre needed to accommodate the observed facts; there seems to be no fundamental reason why symmetry breaking should occur in this way. Although a GUT contains only one gauge coupling constant, there are many other undetermined parameters, such as masses and coupling constants associated with the Higgs fields. Thus, the price of obtaining a prediction for one more measurable quantity, the Weinberg angle, is the introduction of further quantities that cannot even be measured. It would apparently be necessary to conduct experiments at inconceivably high energies to test any specific features of grand unified theories other than proton decay. Finally, grand unification involves a theoretical conundrum known as the *gauge hierarchy problem*. As we saw in §9.6, renormalizability generally requires us to include in the Lagrangian all those terms that are allowed by the symmetries, and do not involve coupling constants of negative dimension. In a grand unified theory, this turns out, in particular, to require interactions between the two sets of Higgs fields, whose vacuum expectation values are v and V. Gauge-boson masses are given by expressions similar to (12.24), and this requires that $v/V \sim M_{\mathrm{W}}/M_{\mathrm{X}} \sim 10^{-13}$. When the Higgs fields interact, the generic outcome of spontaneous symmetry breaking is that $v/V \simeq 1$; the tiny ratio that we need will come about only if the parameters that determine the shape of the potential are very finely tuned so as to make this happen, and this fine tuning seems to demand some explanation.

All in all, the fact that the running coupling constants of the standard model nearly meet at around $10^{15}\,\mathrm{GeV}$ (or, more or less equivalently, that we can obtain a reasonable prediction for $\sin^2\theta_{\mathrm{W}}$) points quite strongly to some kind of underlying grand unification. On the other hand, simply building a bigger and better gauge theory requires too many *ad hoc* assumptions for comfort. A further cause for dissatisfaction with the standard model and its grand-unified generalizations is that the most familiar force of all, namely gravity, is not included. A simple prescription for including gravity would seem to follow from the general considerations of chapter 8. Our fully unified theory should be invariant not only under gauge transformations, but also under general coordinate transformations, and this can be achieved quite straightforwardly by the methods we explored in §7.7. To account for the dynamics of the gravitational fields themselves, we would finally add to our Lagrangian the gravitational action (4.16). As we saw in §7.6.2, small fluctuations in the metric tensor field can be interpreted in terms of spin-2 particles—gravitons—which ought to be the gauge bosons of gravity. Other things being equal, this should provide us with a fully unified quantum theory of all the known forces. Unfortunately, other things are not quite equal. The problem is that the coupling constant for gravity is Newton's constant G which, expressed in natural units, is $G/\hbar c^5 = (1.22 \times 10^{19}\,\mathrm{GeV})^{-2}$.

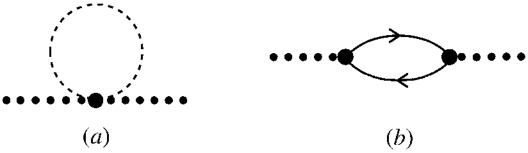


Figure 12.7. Contributions to the self energy of a light scalar particle from (*a*) another scalar and (*b*) a fermion.

According to our discussion in §9.6, the negative dimension of this coupling constant makes the theory non-renormalizable. Remedies for this illness have, naturally, been sought, but none has been found, at least within the context of quantum field theories as we have studied them until now. These difficulties lead many theorists to suspect the existence of some deeper principle.

12.7 Supersymmetry

Part of this deeper principle may be the idea of *supersymmetry*. In general terms, the gauge hierarchy problem might be solved if parameters in the Lagrangian were constrained by a new symmetry in such a way that potentially large contributions, say of order M_X, to the masses of the observed particles would cancel exactly. To set out exactly how this would work needs a more detailed treatment of the inner workings of GUTs than I can give here, but the basic idea is contained in figure 12.7, which shows two contributions to the self-energy of a light particle (the dotted propagators) from a scalar particle (the dashed propagator) and a fermion (the solid propagators). As we know from §9.6, this self-energy represents a correction to the mass of the particle. The key point is that the fermion loop has, as we saw in §9.4, an extra factor of -1 compared with the scalar loop, on account of the anticommutation of the spin-$\frac{1}{2}$ fields. If we could arrange for the magnitudes of these two contributions to be exactly equal, then they would make no net contribution to the mass of the light particle. The symmetry that makes this happen must be one that relates fermions and bosons.

Although the essential idea of supersymmetry is fairly straightforward, a full account of the technology that has been developed to deal with supersymmetric field theories in general might well occupy a book in itself. In this section, I shall illustrate how the symmetry works by studying the simplest example, the Wess–Zumino model, and describe in more qualitative terms how the basic idea might be extended to construct more realistic theories. Much of the literature on supersymmetry uses a special notation for spinors—the van-der-Waerden notation—which I plan to avoid. A detailed introduction to supersymmetry which explains this notation is given by Ryder (1996). A comprehensive account of supersymmetric field theories will be found in Weinberg (2000). Some of the key results require quite tedious algebra, which I shall not always set out in detail. For readers who wish to verify these results, I have collected in §12.7.6 some clues to the manipulations they will find useful.

12.7.1 The Wess–Zumino model

The first obstacle to be overcome in finding a symmetry that relates bosons and fermions is that the two particle species have different spins, and therefore different numbers of spin polarization states available to them. If we want to regard two particles, say A and B, as being (in what will now be a rather esoteric sense) different states of the same basic species, then an A particle and a B particle must have the same number of states available to them. To get this counting of states right, it is helpful in the first instance to deal with massless particles. This is because, as we saw in §7.5, the two helicity states of a massless spin-$\frac{1}{2}$ particle can be treated independently of each other. The supersymmetric model invented by J Wess and B Zumino (1974) contains a single massless fermion. It can be represented by a Majorana spinor $\psi(x)$, for which $\psi^{\mathrm{c}} = C\bar{\psi}^{\mathrm{T}} = \psi$, so that the particle is identical to its antiparticle and can exist in both right-handed and left-handed helicity states. Equivalently, it can be represented by a left-handed spinor ψ_{L}, in which case the particle can exist only in the left-handed state, while its antiparticle can exist only in the right-handed state. The two descriptions are related by

$$\psi_{\mathrm{L}}(x) = P_{\mathrm{L}}\psi(x) \qquad \psi(x) = \psi_{\mathrm{L}}(x) + \psi_{\mathrm{L}}^{\mathrm{c}}(x) = \psi_{\mathrm{L}}(x) + C\bar{\psi}_{\mathrm{L}}^{\mathrm{T}}(x) \quad (12.82)$$

where $P_{\mathrm{L}} = \frac{1}{2}(1-\gamma^5)$ is the projection operator introduced in (7.76). To be explicit about the notation here, $\bar{\psi}_{\mathrm{L}}$ means $\psi_{\mathrm{L}}^{\dagger}\gamma^0 = \bar{\psi}P_{\mathrm{R}}$; this is not the same as $\bar{\psi}P_{\mathrm{L}}$. With either description, the fermion has two independent states, which must be matched by two independent bosonic states. These could be represented either by two real scalar fields or by one complex scalar field. (Later, we shall think about alternatives such as the two helicity states of a massless spin-1 or spin-2 particle). I shall present the model in terms of a left-handed spinor $\psi_L(x)$ and a complex scalar field $\phi(x)$, in which case its Lagrangian density is given by

$$\mathcal{L} = \partial_\mu\phi^*\partial^\mu\phi + \mathrm{i}\bar{\psi}_{\mathrm{L}}\not{\partial}\psi_{\mathrm{L}} + \mathcal{F}^*\mathcal{F}. \quad (12.83)$$

In addition to ϕ and ψ_{L}, it contains a second complex scalar field $\mathcal{F}$, in a form that we have not met before. We can easily see that $\mathcal{F}$ has no real physical meaning, because its Euler–Lagrange equation is $\mathcal{F} = 0$; it is called an *auxiliary field*, and is there to make the mathematics of supersymmetry work smoothly.

An infinitesimal supersymmetry transformation is the change of variables $\phi \to \phi + \delta\phi$, $\psi_{\mathrm{L}} \to \psi_{\mathrm{L}} + \delta\psi_{\mathrm{L}}$, $\mathcal{F} \to \mathcal{F} + \delta\mathcal{F}$, where

$$\delta\phi(x) = \sqrt{2}\bar{\epsilon}\psi_{\mathrm{L}}(x) \quad (12.84)$$

$$\delta\psi_{\mathrm{L}}(x) = -\mathrm{i}\sqrt{2}P_{\mathrm{L}}\gamma^\mu\epsilon\partial_\mu\phi(x) + \sqrt{2}P_{\mathrm{L}}\epsilon\mathcal{F}(x) \quad (12.85)$$

$$\delta\mathcal{F}(x) = -\mathrm{i}\sqrt{2}\bar{\epsilon}\not{\partial}\psi_{\mathrm{L}}(x). \quad (12.86)$$

Clearly, the change in ϕ must be a commuting, scalar quantity, while the change in ψ_{L} must be an anticommuting spinor quantity. Thus, the small parameter ϵ is a constant spinor. That is, it consists of a set of four Grassmann numbers ϵ_α, which transform as a spinor under Lorentz transformations, although they are not field operators. In fact, we take ϵ to be a Majorana spinor, for which $\bar{\epsilon} = \epsilon^{\mathrm{T}} C$. The small changes in the conjugate fields are

$$\delta\phi^*(x) = \sqrt{2}\bar{\psi}_{\mathrm{L}}(x)\epsilon \tag{12.87}$$

$$\delta\bar{\psi}_{\mathrm{L}}(x) = \mathrm{i}\sqrt{2}\partial_\mu\phi^*(x)\bar{\epsilon}\gamma^\mu P_{\mathrm{R}} + \sqrt{2}\mathcal{F}^*(x)\bar{\epsilon}P_{\mathrm{R}} \tag{12.88}$$

$$\delta\mathcal{F}^*(x) = \mathrm{i}\sqrt{2}\bar{\psi}_{\mathrm{L}}(x)\overleftarrow{\not{\partial}}\,\epsilon. \tag{12.89}$$

With the supersymmetry transformation defined in this way, we can work out the small change in the Lagrangian density, keeping only the terms of order ϵ. It is

$$\delta\mathcal{L} = \sqrt{2}\partial_\mu X^\mu(x) \tag{12.90}$$

where

$$X^\mu(x) = \bar{\psi}_{\mathrm{L}}(x)\epsilon\partial^\mu\phi(x) + \tfrac{1}{2}\bar{\epsilon}[\gamma^\mu, \gamma^\nu]\psi_{\mathrm{L}}(x)\partial_\nu\phi^*(x) + \mathrm{i}\bar{\psi}_{\mathrm{L}}\gamma^\mu\epsilon\mathcal{F}. \tag{12.91}$$

Because $\delta\mathcal{L}$ is a total divergence, it does not affect the equations of motion. Its contribution to the action, $\delta S = \sqrt{2}\int \mathrm{d}^4x\ \partial_\mu X^\mu$ can usually be set to zero, given suitable boundary conditions at $|x^\mu| \to \infty$.

Evidently, we have found a symmetry of the Wess–Zumino theory, which relates the bosons and the fermions. However, this is a rather uninteresting theory of massless particles, with no interactions. Moreover, it contains only one supersymmetry multiplet, consisting of a spin-0 and a spin-$\frac{1}{2}$ particle, together with their antiparticles. In the next two subsections, I shall outline how this theory can be extended to incorporate masses and interactions for these particles, and then discuss what other supersymmetry multiplets might exist.

12.7.2 Superfields

Given the somewhat complicated nature of the supersymmetry transformation (12.84)–(12.89), it might seem rather difficult to guess at the terms that can be added to the Lagrangian density without destroying the supersymmetry. Fortunately, a method is available for constructing such terms, which makes use of objects called *superfields*. It will be sufficient for the purposes of our discussion here to regard a superfield simply as a collection of fields (including auxiliary fields) that form a supersymmetry multiplet. For the example we have to hand, the relevant superfield $\Phi(\boldsymbol{\phi}, \boldsymbol{\psi}_{\mathrm{L}}, \boldsymbol{\mathcal{F}})$ is called a *left-chiral superfield*. Its component fields $\boldsymbol{\phi}$, $\boldsymbol{\psi}_{\mathrm{L}}$ and $\boldsymbol{\mathcal{F}}$ are respectively a scalar, a left-handed spinor and another scalar. Under a supersymmetry transformation, they transform according to our previous rules but, as indicated by the boldface notation, they are not necessarily the same as the elementary fields ϕ, ψ_{L} and $\mathcal{F}$ that appear in $\mathcal{L}$. In fact, they generally consist of products of these elementary fields.

The usefulness of superfields lies in the fact that they can be added and multiplied to form new ones. To add two superfields, we simply add their components. Thus, if Φ_1 has the components $(\boldsymbol{\phi}_1, \boldsymbol{\psi}_{1\mathrm{L}}, \boldsymbol{\mathcal{F}}_1)$ and Φ_2 has the components $(\boldsymbol{\phi}_2, \boldsymbol{\psi}_{2\mathrm{L}}, \boldsymbol{\mathcal{F}}_2)$, then $\Phi_1 + \Phi_2$ is the superfield whose components are $(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2, \boldsymbol{\psi}_{1\mathrm{L}} + \boldsymbol{\psi}_{2\mathrm{L}}, \boldsymbol{\mathcal{F}}_1 + \boldsymbol{\mathcal{F}}_2)$. It is easy to check that these new components transform in the right way to be a supersymmetry multiplet. To multiply superfields correctly needs a little more care. Let us denote the product Φ of two superfields Φ_1 and Φ_2 by $\Phi = \Phi_1 \circ \Phi_2$. To make this meaningful, we need a rule for constructing the components $(\boldsymbol{\phi}, \boldsymbol{\psi}_{\mathrm{L}}, \boldsymbol{\mathcal{F}})$ of Φ from those of Φ_1 and Φ_2. The rule is

$$\boldsymbol{\phi} = \boldsymbol{\phi}_1\, \boldsymbol{\phi}_2 \tag{12.92}$$

$$\boldsymbol{\psi}_{\mathrm{L}} = \boldsymbol{\phi}_1\, \boldsymbol{\psi}_{2\mathrm{L}} + \boldsymbol{\phi}_2\, \boldsymbol{\psi}_{1\mathrm{L}} \tag{12.93}$$

$$\boldsymbol{\mathcal{F}} = \boldsymbol{\phi}_1\, \boldsymbol{\mathcal{F}}_2 + \boldsymbol{\phi}_2\, \boldsymbol{\mathcal{F}}_1 - \boldsymbol{\psi}_{1\mathrm{L}}^{\mathrm{T}} C \boldsymbol{\psi}_{2\mathrm{L}}. \tag{12.94}$$

If Φ is to be a valid superfield, then $\boldsymbol{\phi}$, $\boldsymbol{\psi}_{\mathrm{L}}$ and $\boldsymbol{\mathcal{F}}$ must have the correct supersymmetry transformations, and it is not too hard to check that they do. It is also not hard to check that $\Phi_1 \circ \Phi_2 = \Phi_2 \circ \Phi_1$, so the order of the superfields does not matter. As I have presented it, this definition of the superfield product is a guess that turns out to work. There is, though, a more general formalism, within which it arises quite naturally. According to this formalism (which I shall not develop in detail), the superfields inhabit an 8-dimensional ‘spacetime’, called *superspace*. The extra four coordinates θ^{α} are Grassmann variables. They have no physical meaning that I know of, but they constitute a useful bookkeeping device.

We can use the superfield idea in the following way to add new supersymmetric terms to the Lagrangian density (12.83). The criterion is that, under an infinitesimal supersymmetry transformation, the new terms must change only by a total divergence. Because the transformation (12.84)–(12.86) applies to *any* superfield, we see that the $\boldsymbol{\mathcal{F}}$ component of any superfield changes by a total divergence, and will suit our purpose. From now on, Φ will stand for our multiplet of elementary fields $(\phi(x), \psi_{\mathrm{L}}(x), \mathcal{F}(x))$. From it, we construct a new superfield, called the *superpotential*,

$$W(\Phi) = \tfrac{1}{2} m \Phi \circ \Phi + \tfrac{1}{6} g \Phi \circ \Phi \circ \Phi \tag{12.95}$$

where m and g are constants. So far as supersymmetry is concerned, we might include higher powers of Φ as well, but these would lead to a non-renormalizable theory. The $\boldsymbol{\mathcal{F}}$ component of the superpotential is

$$W(\Phi)|_{\mathcal{F}} = \tfrac{1}{2} m (2\phi\mathcal{F} - \bar{\psi}_{\mathrm{R}} \psi_{\mathrm{L}}) + \tfrac{1}{2} g (\phi^2 \mathcal{F} - \phi \bar{\psi}_{\mathrm{R}} \psi_{\mathrm{L}}) \tag{12.96}$$

where I have used the fact that $\psi_{\mathrm{L}}^{\mathrm{T}} C \psi_{\mathrm{L}} = \bar{\psi}_{\mathrm{R}} \psi_{\mathrm{L}}$ (see (12.142)). The Lagrangian density must be real, so we add to (12.83) the combination $W(\Phi)|_{\mathcal{F}} +$

$\left[W(\Phi)|_{\mathcal{F}}\right]^*$. Using the fact that $(\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}})^* = \bar\psi_{\mathrm{L}}\psi_{\mathrm{R}}$, we get

$$\begin{aligned}\mathcal{L} = {} & \partial_\mu\phi^*\partial^\mu\phi + \mathrm{i}\bar\psi_{\mathrm{L}}\partial\!\!\!/\psi_{\mathrm{L}} + \mathcal{F}^*\mathcal{F} \\ & + m\left[\phi\mathcal{F} + \phi^*\mathcal{F}^* - \tfrac{1}{2}(\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}} + \bar\psi_{\mathrm{L}}\psi_{\mathrm{R}})\right] \\ & + \tfrac{1}{2}g\left[\phi^2\mathcal{F} + \phi^{*2}\mathcal{F}^* - (\phi\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}} + \phi^*\bar\psi_{\mathrm{L}}\psi_{\mathrm{R}})\right]\end{aligned} \tag{12.97}$$

and this can be rewritten as

$$\begin{aligned}\mathcal{L} = {} & \partial_\mu\phi^*\partial^\mu\phi - m^2\phi^*\phi + \mathrm{i}\bar\psi_{\mathrm{L}}\partial\!\!\!/\psi_{\mathrm{L}} - \tfrac{1}{2}m(\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}} + \bar\psi_{\mathrm{L}}\psi_{\mathrm{R}}) + \widetilde{\mathcal{F}}^*\widetilde{\mathcal{F}} \\ & - \tfrac{1}{2}mg\phi^*\phi(\phi + \phi^*) - \tfrac{1}{4}g^2(\phi^*\phi)^2 - \tfrac{1}{2}g(\phi\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}} + \phi^*\bar\psi_{\mathrm{L}}\psi_{\mathrm{R}})\end{aligned} \tag{12.98}$$

where the new auxiliary field is $\widetilde{\mathcal{F}} = \mathcal{F} + m\phi^* + \frac{1}{2}g\phi^{*2}$. Again, the Euler–Lagrange equation $\widetilde{\mathcal{F}} = 0$ means that we can ignore $\widetilde{\mathcal{F}}$ for practical purposes. In terms of the Majorana field $\psi = \psi_{\mathrm{L}} + \psi_{\mathrm{R}}$, we can write

$$\mathrm{i}\bar\psi_{\mathrm{L}}\partial\!\!\!/\psi_{\mathrm{L}} - \tfrac{1}{2}m(\bar\psi_{\mathrm{R}}\psi_{\mathrm{L}} + \bar\psi_{\mathrm{L}}\psi_{\mathrm{R}}) = \tfrac{1}{2}\bar\psi(\mathrm{i}\partial\!\!\!/ - m)\psi \tag{12.99}$$

up to a total divergence, which can also be ignored.

As we might have expected, this supersymmetric theory describes spin-0 and spin-$\frac{1}{2}$ particles that have exactly the same mass, m. In fact it can be shown (though I shall not prove it here) that there is no mass renormalization in this theory. That is to say, the mass parameter m is actually the physical mass of the particles; the corrections that are potentially present at any order of perturbation theory are guaranteed to cancel. This happens because the various interaction terms in (12.98) have coupling constants that are related in a special way, in order to make the theory supersymmetric. This *nonrenormalization* property is just the sort of feature that might alleviate the gauge hierarchy problem, if it could be incorporated into a grand unified theory.

12.7.3 Spontaneous supersymmetry breaking

In nature, there are no known examples of bosons and fermions having identical masses. For this reason alone (there is another that we shall meet a little later on), supersymmetry cannot be a feature of the world as we know it. There are two ways in which supersymmetry might nevertheless be relevant at a fundamental level. One is that it might be spontaneously broken, just as the gauge symmetry of the standard model or of a grand unified theory is. The other is that supersymmetry might only be approximately true, even at a fundamental level, in which case it is said to be 'explicitly' broken. The latter possibility is not an attractive one, because it is likely to spoil the exact cancellations which are the principal advantage of having a supersymmetric theory in the first place. On the other hand, spontaneous breaking of supersymmetry does not occur as readily as the spontaneous breaking of other symmetries. In fact, it is impossible in the Wess–Zumino model. To see why, it is helpful to rewrite the Lagrangian density (12.98) in terms of the superpotential. Let us in fact consider a more general model, which will be useful shortly, containing several left-chiral supermultiplets. Its superpotential $W(\Phi_1, \ldots, \Phi_n)$ is a cubic polynomial in the various superfields, and the Lagrangian density can be expressed (leaving out the auxiliary fields $\widetilde{\mathcal{F}}_i$) as

$$\mathcal{L} = \mathcal{L}_0 - V(\phi_1, \ldots, \phi_n) - \frac{1}{2} \sum_{i,j=1}^{n} \left[\frac{\partial^2 W}{\partial\phi_i \partial\phi_j} \bar{\psi}_{i\mathrm{R}} \psi_{j\mathrm{L}} + \left(\frac{\partial^2 W}{\partial\phi_i \partial\phi_j} \right)^* \bar{\psi}_{i\mathrm{L}} \psi_{j\mathrm{R}} \right] \tag{12.100}$$

where $\mathcal{L}_0$ is a sum of terms of the form $\partial_\mu \phi_i^* \partial^\mu \phi_i$ and $\mathrm{i}\bar{\psi}_{i\mathrm{L}} \not{\partial} \psi_{i\mathrm{L}}$. Here, the superpotential $W(\phi_1, \ldots, \phi_n)$ is now an ordinary function just of the scalar components of the multiplets, and the potential is

$$V(\phi_1, \ldots, \phi_n) = \sum_{i=1}^{n} \left| \frac{\partial W}{\partial \phi_i} \right|^2 . \tag{12.101}$$

Readers should find this easy to verify for the case of a single multiplet, and it follows more generally (though less obviously) from the procedure for constructing $W|_{\mathcal{F}}$. For a single multiplet, suppose that ϕ acquires a vacuum expectation value v. Writing $\phi(x) = v + \widetilde{\phi}(x)$, we can expand the potential as

$$V(\phi) = V(v) + [W''(v)]^2 \widetilde{\phi}^* \widetilde{\phi} + \ldots \tag{12.102}$$

which shows that both the scalar particle and the spin-$\frac{1}{2}$ particle have the same mass, $m = W''(v)$, regardless of the value of v.

A simple criterion that shows what is needed for supersymmetry to be spontaneously broken can be found from the transformation (12.84)–(12.86). This transformation must produce some change in the vacuum state, analogous to moving around the circle of minima in the potential of figure 11.8. Therefore, the vacuum expectation value of at least one of the small changes must be different from zero. Only a scalar field can have a non-zero expectation value (otherwise, the vacuum would have a non-zero angular momentum) and we assume that the vacuum is homogeneous, so that $\langle 0|\partial_\mu \phi(x)|0\rangle = 0$. The only possibility is that $\langle 0|\widetilde{\mathcal{F}}|0\rangle \neq 0$. But in (12.98), the Euler–Lagrange equation tells us that $\widetilde{\mathcal{F}} = \widetilde{\mathcal{F}}^* = 0$, and this implies that $W'(v) = -\langle 0|\mathcal{F}|0\rangle \neq 0$. In this way, we discover that supersymmetry will be spontaneously broken only if $V(v) = \left|W'(v)\right|^2 > 0$. The same criterion holds for the more general potential (12.101). Now, V is a sum of positive quantities and cannot be negative. If there is some set of values of the fields for which $V = 0$, then this will be a minimum and supersymmetry will be unbroken. If supersymmetry is to be spontaneously broken, then V must be a function that does not vanish for any values of the ϕ_i.

Possibly the simplest model that does exhibit spontaneous supersymmetry breaking is one invented by L O'Raifeartaigh, which contains three left-chiral multiplets whose scalar components are, say, ϕ, χ_1 and χ_2. The superpotential for this model is

$$W(\phi, \chi_1, \chi_2) = m\chi_1\phi + \tfrac{1}{2}g\chi_2(\phi^2 - \lambda) \tag{12.103}$$

where m, g and λ are constants; it has the crucial feature that two functions of ϕ, namely ϕ and $\phi^2 - \lambda$, which cannot both vanish at the same time, appear multiplied by the independent fields χ_1 and χ_2. From (12.101), we derive the potential

$$V = |m\chi_1 + g\chi_2\phi|^2 + m^2|\phi|^2 + \tfrac{1}{4}g^2|\phi^2 - \lambda|^2 \tag{12.104}$$

whose first term can be minimized, without affecting the minimization of the remaining terms, by taking $\chi_1 = \chi_2 = 0$. Clearly, the two other terms cannot both vanish, so supersymmetry is spontaneously broken. To find the minimum, we must solve the equation

$$\frac{\partial V}{\partial \phi} = m^2\phi^* + \tfrac{1}{2}g^2\phi(\phi^{*2} - \lambda) = 0 \tag{12.105}$$

and its complex conjugate. If $m^2 > \frac{1}{2}g^2|\lambda|$ (as I shall assume to make things simple), then the only solution is $\phi = 0$. We find the masses of the scalar particles by expanding V about the minimum $\phi = \chi_1 = \chi_2 = 0$. The terms quadratic in the fields are

$$m^2\chi_1^*\chi_1 + \tfrac{1}{2}(m^2 - \tfrac{1}{2}\lambda g^2)\phi_1^2 + \tfrac{1}{2}(m^2 + \tfrac{1}{2}\lambda g^2)\phi_2^2 \tag{12.106}$$

where I have written the complex field ϕ in terms of its real and imaginary parts as $\phi = (\phi_1 + \mathrm{i}\phi_2)/\sqrt{2}$. The particles and antiparticles associated with the complex field χ_1 have a mass m, while those associated with χ_2 are massless. The two particles associated with the real fields ϕ_1 and ϕ_2 have masses equal to $(m^2 \pm \frac{1}{2}\lambda g^2)^{1/2}$ and each one is its own antiparticle. (As in the electroweak theory, the field operators that create and annihilate physical particles, with definite masses, are those linear combinations of the original set of fields that diagonalize the quadratic terms in $\mathcal{L}$). What about the masses of the fermions? To find these, we have to evaluate the matrix that multiplies the fermionic term in (12.100) at the field values that minimize the potential. That is

$$\left.\frac{\partial^2 W}{\partial\phi_i\partial\phi_j}\right|_{\phi=\chi_1=\chi_2=0} = \begin{pmatrix} 0 & m & 0 \\ m & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{12.107}$$

Again, the field operators for physical particles are the linear combinations that diagonalize this 'mass matrix' and the particle masses are the eigenvalues, namely m, m and 0. Although there are both bosons and fermions of mass m, and both bosons and fermions with vanishing mass, the masses of all the bosons do not match those of all the fermions, so supersymmetry is indeed broken. The appearance of a massless spin-$\frac{1}{2}$ fermion is a general feature of spontaneous supersymmetry breaking, quite analogous to the Goldstone boson we encountered in §11.7.1. This massless particle is called the *Goldstone fermion*, or more often (I regret to say) the 'Goldstino'.

12.7.4 The supersymmetry algebra

If we want to incorporate supersymmetry into our gauge theories of fundamental interactions, we need to identify supersymmetry multiplets of particles that contain spin-1 particles (and, indeed, spin-2 particles if we hope to include gravity). It will be useful, then, to know just what supersymmetry multiplets are possible. The tool we need to find this out is a structure analogous to the commutation relations (8.28) satisfied by the generators of a symmetry group such as SU(2). We shall find it helpful to cast this structure in terms of operators that act in the Hilbert space of state vectors; in fact, we could well have dealt with gauge symmetries in this way, but we should not have gained much by doing so. Consider the definition (5.35) of a Heisenberg-picture operator, which we can rewrite as

$$A(t) = \mathrm{e}^{\mathrm{i}tH} A \mathrm{e}^{-\mathrm{i}tH}. \tag{12.108}$$

The Hamiltonian H is, of course, the generator of time translations, and it should be obvious that we can shift the time to which $A(t)$ refers by an amount a^0 (which will shortly become one component of a 4-vector) by using the time evolution operator

$$A(t + a^0) = \mathrm{e}^{\mathrm{i}a^0 H} A(t) \mathrm{e}^{-\mathrm{i}a^0 H}. \tag{12.109}$$

In a relativistic theory, we can generate translations of a field operator $A(x)$ through a 4-vector a^μ in the same way. That is, we can write

$$A(x + a) = \mathrm{e}^{\mathrm{i}a\cdot P} A(x) \mathrm{e}^{-\mathrm{i}a\cdot P}. \tag{12.110}$$

where $P^0 = H$ is the Hamiltonian, and the spatial components P^i are the operators corresponding to the total linear momentum of our system, which are the generators of space translations. In principle, we could construct expressions for the P^i, and for other symmetry generators that we shall meet shortly, in terms of field operators, as we did in (7.21) for the Hamiltonian, but it will not generally be necessary to do this explicitly. For an infinitesimal translation, we can find the small change in $A(x)$ by expanding both sides of (12.110) in powers of a^μ. We get

$$A(x) + a^\mu \partial_\mu A(x) + \ldots = A(x) + \mathrm{i}a^\mu P_\mu A(x) - A(x)\mathrm{i}a^\mu P_\mu + \ldots \tag{12.111}$$

so we can identify

$$\delta_a A(x) = a^\mu \partial_\mu A(x) = \mathrm{i}[a^\mu P_\mu, A(x)]. \tag{12.112}$$

The first expression gives the change in $A(x)$ explicitly, while the second one indicates how this change is produced by the operators P_μ.

The Lorentz transformations and rotations that we studied in §7.3.2 can be dealt with in the same way. For a Dirac spinor, we have

$$\begin{aligned}\exp(\tfrac{1}{2}\mathrm{i}\omega_{\mu\nu}M^{\mu\nu})\psi(x)\exp(-\tfrac{1}{2}\mathrm{i}\omega_{\mu\nu}M^{\mu\nu})&\\ &= \exp(-\tfrac{1}{2}\mathrm{i}\omega_{\mu\nu}m^{\mu\nu})\psi(x)\\ &= (I-\tfrac{1}{2}\mathrm{i}\omega_{\mu\nu}m^{\mu\nu}+\ldots)\psi(x). \qquad (12.113)\end{aligned}$$

In this case, $M^{\mu\nu}$ are the Hilbert-space operators corresponding to the generators of Lorentz transformations, and $m^{\mu\nu}$ are the combined matrix and differential operators that I previously denoted by $M^{\mu\nu}$ in (7.41). The infinitesimal change in a general field operator $A(x)$ will be given by

$$\delta_\omega A(x) = \tfrac{1}{2}\omega_{\mu\nu}\left(-\mathrm{i}\Sigma^{\mu\nu}+x^\mu\partial^\nu - x^\nu\partial^\mu\right)A(x) = \tfrac{1}{2}\mathrm{i}[\omega_{\mu\nu}M^{\mu\nu}, A(x)] \qquad (12.114)$$

where $\Sigma^{\mu\nu}$ is the spin matrix appropriate for $A(x)$; for example, $\Sigma^{\mu\nu}=0$ if $A(x)$ is a scalar field and $\Sigma^{\mu\nu}=\frac{1}{2}\sigma^{\mu\nu}=\frac{1}{4}\mathrm{i}[\gamma^\mu,\gamma^\nu]$ if $A(x)$ is a spinor.

The supersymmetry transformations we have been discussing have four generators, Q_α, which are the four components of a Majorana spinor, in the same way that the operators P^μ are the components of a 4-vector. A transformed field is given by

$$A'(x) = \mathrm{e}^{\mathrm{i}\bar{\epsilon}Q}A(x)\mathrm{e}^{-\mathrm{i}\bar{\epsilon}Q} \qquad (12.115)$$

in which $\bar{\epsilon}Q = \epsilon^{\mathrm{T}}CQ = C_{\alpha\beta}\epsilon_\alpha Q_\beta$ is a Lorentz-invariant quantity. (According to a commonly-used terminology, the generators of symmetry transformations are called 'charges' and, in particular, the Q_α are 'supercharges'. There is clearly an analogy with expressions such as (8.53), where Q is electric charge. According to Noether's theorem, these 'charges' are conserved quantities when our theory is invariant under the corresponding symmetries.) For an infinitesimal transformation, we have

$$\delta_\epsilon A(x) = \mathrm{i}[\bar{\epsilon}Q, A(x)] \qquad (12.116)$$

and, for the fields in our left-chiral supermultiplet, the explicit expressions for the small changes $\delta_\epsilon A(x)$ are those given in (12.84)–(12.86).

The key information to be extracted from this operator formalism is the commutation relations enjoyed by the generators Q_α. They can be determined by asking about the effect of two successive infinitesimal transformations. Suppose that we have made one transformation using parameters ϵ_1, leading to small changes $\delta_{\epsilon_1}A(x)$. Now we ask about the change in $\delta_{\epsilon_1}A(x)$ upon making a second transformation, with parameters ϵ_2. It is given by

$$\begin{aligned}\delta_{\epsilon_2}(\delta_{\epsilon_1}A(x)) &= \mathrm{i}[\bar{\epsilon}_2 Q, \delta_{\epsilon_1}A(x)]\\ &= \mathrm{i}[\bar{\epsilon}_2 Q, \mathrm{i}[\bar{\epsilon}_1 Q, A(x)]]\end{aligned}$$

$$\begin{aligned} = & -\bar{\epsilon}_2 Q\bar{\epsilon}_1 QA(x) - A(x)\bar{\epsilon}_1 Q\bar{\epsilon}_2 Q \\ & + \bar{\epsilon}_1 QA(x)\bar{\epsilon}_2 Q + \bar{\epsilon}_2 QA(x)\bar{\epsilon}_1 Q. \end{aligned} \qquad (12.117)$$

On comparing this with the result of making the two transformations in the reversed order, we find

$$\delta_{\epsilon_1}(\delta_{\epsilon_2}A(x)) - \delta_{\epsilon_2}(\delta_{\epsilon_1}A(x)) = \mathrm{i}[\mathrm{i}[\bar{\epsilon}_1 Q, \bar{\epsilon}_2 Q], A(x)]. \qquad (12.118)$$

By calculating the left-hand side from our explicit expressions for the changes in the fields, we can identify the operator that is equal to $[\bar{\epsilon}_1 Q, \bar{\epsilon}_2 Q]$. Let us do this, taking $A(x)$ to be the scalar field $\phi(x)$. The individual terms are

$$\begin{aligned} \delta_{\epsilon_1}(\delta_{\epsilon_2}\phi(x)) &= \delta_{\epsilon_1}(\sqrt{2}\bar{\epsilon}_2\psi_{\mathrm{L}}(x)) \\ &= 2[\mathrm{i}\bar{\epsilon}_2 P_{\mathrm{L}}\gamma^\mu\epsilon_1\partial_\mu\phi(x) + \bar{\epsilon}_2 P_{\mathrm{L}}\epsilon_1\mathcal{F}(x)] \\ &= 2[-\mathrm{i}\bar{\epsilon}_1 P_{\mathrm{R}}\gamma^\mu\epsilon_2\partial_\mu\phi(x) + \bar{\epsilon}_1 P_{\mathrm{L}}\epsilon_2\mathcal{F}(x)] \qquad (12.119) \\ \delta_{\epsilon_2}(\delta_{\epsilon_1}\phi(x)) &= 2[\mathrm{i}\bar{\epsilon}_1 P_{\mathrm{L}}\gamma^\mu\epsilon_2\partial_\mu\phi(x) + \bar{\epsilon}_1 P_{\mathrm{L}}\epsilon_2\mathcal{F}(x)] \qquad (12.120) \end{aligned}$$

and by subtracting these two results we deduce

$$\begin{aligned} \delta_{\epsilon_1}(\delta_{\epsilon_2}\phi(x)) - \delta_{\epsilon_2}(\delta_{\epsilon_1}\phi(x)) = & -2\mathrm{i}\bar{\epsilon}_1\gamma^\mu\epsilon_2\partial_\mu\phi(x) \\ = & \, 2\mathrm{i}\bar{\epsilon}_1^\alpha\bar{\epsilon}_2^\beta(\gamma^\mu C)_{\alpha\beta}\partial_\mu\phi(x) \\ = & -2\bar{\epsilon}_1^\alpha\bar{\epsilon}_2^\beta(\gamma^\mu C)_{\alpha\beta}[P_\mu, \phi(x)]. \end{aligned} \qquad (12.121)$$

Comparing our result with (12.118), we find

$$[\bar{\epsilon}_1 Q, \bar{\epsilon}_2 Q] = 2\bar{\epsilon}_1^\alpha\bar{\epsilon}_2^\beta(\gamma^\mu C)_{\alpha\beta}P_\mu. \qquad (12.122)$$

Strictly speaking, we have found out only how these operators act on $\phi(x)$. To make sure that (12.122) is a valid relation between the operators Q_α and P_μ, we should check that the same result comes from acting on $\psi_{\mathrm{L}}(x)$ or on $\mathcal{F}(x)$, and energetic readers may like to do this.

We have not quite reached our result, because (12.122) still contains the parameters ϵ_α. Remembering that these anticommute with each other and with fermionic operators such as Q_α, we have

$$[\bar{\epsilon}_1 Q, \bar{\epsilon}_2 Q] = \bar{\epsilon}_1^\alpha Q_\alpha\bar{\epsilon}_2^\beta Q_\beta - \bar{\epsilon}_2^\beta Q_\beta\bar{\epsilon}_1^\alpha Q_\alpha = -\bar{\epsilon}_1^\alpha\bar{\epsilon}_2^\beta\{Q_\alpha, Q_\beta\} \qquad (12.123)$$

where, as with the creation and annihilation operators for spin-$\frac{1}{2}$ particles, $\{Q_\alpha, Q_\beta\}$ is the anticommutator. Thus, we finally find

$$\{Q_\alpha, Q_\beta\} = -2(\gamma^\mu C)_{\alpha\beta}P_\mu. \qquad (12.124)$$

This can also be written as

$$\{Q_\alpha, \bar{Q}_\beta\} = 2(\gamma^\mu)_{\alpha\beta}P_\mu \qquad (12.125)$$

because $\bar{Q}_\beta = (Q^{\mathrm{T}}C)_\beta$ and $C^2 = -1$. Compare this result with (8.28). We see first that the elements of the matrices γ^μ or $\gamma^\mu C$ serve as the structure constants C^{abc} for the supersymmetry algebra. It is also apparent that the supersymmetry generators Q_α alone do not form a structure analogous to the Lie algebra of one of our earlier symmetry groups, because the momentum operator appears on the right-hand side of (12.124) and must also be considered as a part of this structure. For this reason, we should include the commutators $[P_\mu, P_\nu]$ and $[Q_\alpha, P_\mu]$. By the method we have just used to find (12.124), it is easy to show that

$$[P_\mu, P_\nu] = [Q_\alpha, P_\mu] = 0. \tag{12.126}$$

Thus, the generators of supersymmetry *and* spacetime translations together form a complete structure (or, in the usual terminology, a 'closed' structure). This structure is not quite the same as a Lie algebra of the kind that we met in chapter 8, because it involves both commutators and anticommutators. It is called a *graded Lie algebra*. The graded Lie algebra that we have found is usually considered as part of a larger one, which also includes the generators $M^{\mu\nu}$ of Lorentz transformations and rotations, and is called the *super-Poincaré algebra*. For our present purposes, we need to know only the commutator

$$[Q_\alpha, M^{\mu\nu}] = \tfrac{1}{2}(\sigma^{\mu\nu})_{\alpha\beta}Q_\beta \tag{12.127}$$

which can be established by the same method as before. We have derived the set of (anti)commutation relations (12.124)–(12.127) from the field transformations that we already knew to constitute a symmetry of the Wess–Zumino model. To progress, we must now suppose that the same graded Lie algebra will apply to other theories that exhibit supersymmetry. As a matter of fact, more general versions of the supersymmetry idea (called *extended supersymmetries*) can be constructed, but I shall not deal with them here.

After this rather lengthy preamble, we are ready to address the question of what species of particles can be grouped into supersymmetry multiplets. First, consider the fact that the supercharges Q_α form a Majorana spinor. This means that $Q^{\mathrm{c}} = C\bar{Q}^{\mathrm{T}}$ is equal to Q, or $(C\gamma^{0\mathrm{T}})_{\alpha\beta}Q^\dagger_\beta = Q_\alpha$. Using the Weyl representation of the γ matrices given in §7.5, readers should find it an easy matter to work out the elements of the matrix $C\gamma^{0\mathrm{T}}$ and hence to verify that

$$Q_3 = -Q_2^\dagger \qquad \text{and} \qquad Q_4 = Q_1^\dagger. \tag{12.128}$$

We shall concentrate on the supermultiplets that can be formed from massless particles; masses can be introduced at a later stage as we saw in §12.7.2. In particular, consider a single-particle state, in which the particle has 4-momentum $p^\mu = (p, 0, 0, p)$. Because of Lorentz covariance, our deductions about this state will apply equally to states with other 4-momenta. Since Q_α commutes with P^μ, the result of acting with Q_α on such a state will be another state with the same 4-momentum, so we can consider just the subspace of the whole Hilbert space that consists of all these states. Within this subspace, the operator P^μ can be replaced with the eigenvalues p^μ. In particular, the anticommutation relation (12.124) becomes

$$\{Q_\alpha, Q_\beta\} = -2p(\gamma^0 C - \gamma^3 C)_{\alpha\beta} = 4p \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}_{\alpha\beta} . \qquad (12.129)$$

Taking this together with (12.128), the anticommutation relations for the Q_α can be summarized as

$$Q_1 Q_1^\dagger + Q_1^\dagger Q_1 = 4p \qquad (12.130)$$

$$Q_2 Q_2^\dagger + Q_2^\dagger Q_2 = 0 \qquad (12.131)$$

$$Q_1 Q_1 = Q_1^\dagger Q_1^\dagger = Q_2 Q_2 = Q_2^\dagger Q_2^\dagger = 0. \qquad (12.132)$$

The second one, (12.131), allows us to dispense with Q_2 altogether, as can be seen in the following way. Let $|\Psi\rangle$ be one of the states in our subspace, and let $|\Psi'\rangle = Q_2|\Psi\rangle$ and $|\Psi''\rangle = Q_2^\dagger|\Psi\rangle$, which implies $\langle\Psi'| = \langle\Psi|Q_2^\dagger$ and $\langle\Psi''| = \langle\Psi|Q_2$. Then we find from (12.131) that

$$\langle\Psi|\left(Q_2 Q_2^\dagger + Q_2^\dagger Q_2\right)|\Psi\rangle = \langle\Psi''|\Psi''\rangle + \langle\Psi'|\Psi'\rangle = 0. \qquad (12.133)$$

But neither $\langle\Psi''|\Psi''\rangle$ nor $\langle\Psi'|\Psi'\rangle$ can be negative, so they must both vanish. This means that Q_2 and $Q_2^\dagger$ give zero when acting on any vector in the subspace and can be ignored.

As we know from §§7.5 and 7.6, massless particles of spin s can exist only in states of definite helicity, with spin components of $\pm s$ in the direction of their 3-momenta. Within our subspace, the relevant component of angular momentum is $J^3 = M^{12}$. The spin matrix on the right-hand side of the commutation relation (12.127) is easily found to be

$$\tfrac{1}{2}\sigma^{12} = \tfrac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \qquad (12.134)$$

so we finally discover the commutators that will yield our answers:

$$[Q_1, J^3] = \tfrac{1}{2} Q_1 \qquad \text{and} \qquad [Q_1^\dagger, J^3] = -\tfrac{1}{2} Q_1^\dagger. \qquad (12.135)$$

Comparing these with the commutators of the energy raising and lowering operators (5.60) and (5.61) of the harmonic oscillator, we see that acting with Q_1 reduces the helicity J^3 of a state by $\frac{1}{2}$, while acting with $Q_1^\dagger$ increases the helicity by $\frac{1}{2}$. In fact, we can repeat the argument that gave us the energy spectrum of the harmonic oscillator to find all the allowed helicity values in a supermultiplet. Within the multiplet, let $|h_{\rm min}\rangle$ be the state of lowest helicity, for which $Q_1|h_{\rm min}\rangle = 0$. Acting with $Q_1^\dagger$ on this state, we get a state of helicity $h_{\rm min} + \frac{1}{2}$,

$$Q_1^\dagger|h_{\rm min}\rangle = |h_{\rm min} + \tfrac{1}{2}\rangle. \qquad (12.136)$$

Acting with $Q_1^\dagger$ on $|h_{\rm min} + \frac{1}{2}\rangle$ gives zero, on account of (12.132), while acting with Q_1 gives us back our original state:

$$Q_1|h_{\rm min} + \tfrac{1}{2}\rangle = Q_1 Q_1^\dagger|h_{\rm min}\rangle = (4p - Q_1^\dagger Q_1)|h_{\rm min}\rangle = 4p|h_{\rm min}\rangle. \qquad (12.137)$$

We see, then, that each supersymmetry multiplet consists of just two states, say with helicity h and $h + \frac{1}{2}$. As in the Wess–Zumino model (for which $h = 0$), a supersymmetric theory containing this multiplet will also contain the multiplet of antiparticles, with helicities $-h$ and $-(h + \frac{1}{2})$. In terms of particles, a supermultiplet contains just two particles—one with spin s and one with spin $s + \frac{1}{2}$.

12.7.5 Supersymmetric gauge theories and supergravity

The business of constructing supersymmetric gauge theories that might have some relevance to the real world is rather too complicated for me to give any detailed account here, but some general features can be appreciated without too much trouble. It is easy to see, for example, that the standard model is not supersymmetric. In a supersymmetric theory, each of the gauge bosons would belong to a supermultiplet, with a partner whose spin is either $s = \frac{1}{2}$ or $s = \frac{3}{2}$. No fundamental spin-$\frac{3}{2}$ particles are known, so let us suppose that the partners are spin-$\frac{1}{2}$ particles. Then if weak isospin and supersymmetry are both to be valid symmetries, each weak-isospin multiplet must itself be composed of supermultiplets. Since the $\mathrm{W}^\pm$ bosons have isospin $t = 1$, their spin-$\frac{1}{2}$ partners should also have $t = 1$. They cannot, therefore, be identified with the quarks or leptons, whose left-handed components have $t = \frac{1}{2}$ and whose right-handed components have $t = 0$. In fact, if the world is supersymmetric, then all of the known particles must have distinct superpartners. According to the traditional terminology for these 'sparticles', there would exist scalar partners for the quarks and leptons (the 'squarks' and 'sleptons'), spin-$\frac{1}{2}$ partners for the gauge bosons (the 'wino', 'zino', 'photino' and 'gluinos') and a spin-$\frac{1}{2}$ partner for Higgs particle (whose name I leave it as an exercise for readers to determine). More accurately, it turns out that two multiplets of Higgs fields are needed to generate all the fermion masses within a supersymmetric theory.

Here, then, is the second reason why supersymmetry must be broken. Not only do we observe no pairs of bosons and fermions with the same masses, but no particle observed to date can be identified as the supersymmetric partner of any other known particle. If supersymmetry has anything to do with the real world, it must be spontaneously broken in such a way that all the partners of the known particles have masses that are too large for these ‘sparticles’ to be produced at currently accessible energies. Mechanisms that achieve this can be invented, but there is no consensus on which of them, if any, might be correct. Indeed, the whole idea might seem to be an unpromising, *ad hoc* contrivance, were it not for one intriguing piece of evidence. A theoretical model has been constructed, which is called the *minimal supersymmetric standard model* (or MSSM), although it is not, strictly speaking, supersymmetric. Rather, it is to be thought of as what remains of an underlying, genuinely supersymmetric (but largely unspecified) theory, when the fields associated with particles that acquire very large masses through spontaneous symmetry breaking are left out. The fields in the MSSM are those for the known particles and their superpartners, but its Lagrangian density contains terms that break supersymmetry explicitly. By adjusting its parameters appropriately, this model can be made to agree with experimental data as accurately as the ordinary standard model, but it achieves one more success. This comes about because the existence of extra particles changes the numbers β_i in (12.74)–(12.76) that determine the variation of the running coupling constants with energy. As I mentioned in §12.6, the coupling constants calculated from the standard model become very similar at a unification energy of about 10^{15} GeV, but they do not all become equal at exactly the same energy. When this calculation is repeated, using the β_i of the MSSM, it is found that they do all become equal (within the accuracy of experimental data, which are now quite precise), at an energy of about 2×10^{16} GeV. At the time of writing, this is the one indication from actual observations that the mathematics of supersymmetry may be relevant to particle physics. (It is worth mentioning, though, that some of the mathematics of supersymmetry has found applications to certain problems in condensed-matter physics, which mainly concern disordered systems. Interested readers may like to consult the book by Efetov (1997).)

We saw in §12.7.3 that the spontaneous breaking of supersymmetry implies the existence of a massless Goldstone fermion. Even if all the other ‘sparticles’ have masses that are too large for them to have been observed, this cannot apply to the Goldstone fermion, and there is no known particle with which it can be identified. However, the Goldstone bosons that might be expected from the spontaneous breaking of the electroweak symmetry are also not observed. The reason, as I explained in §11.7.3, is that when a *local* gauge symmetry is spontaneously broken, the ‘would-be Goldstone boson’ appears not as a massless spin-0 particle, but rather as the zero-helicity component of the massive gauge boson. We might wonder, then, whether supersymmetry can be promoted to some kind of local gauge symmetry, so that the theory would be invariant under transformations similar to (12.84)–(12.86), but with a spacetime-dependent parameter $\epsilon(x)$. This is indeed possible, though algebraically too complicated for me to give more than some qualitative remarks. It is crucial to remember that supersymmetry transformations on their own do not form a closed algebraic structure. As we saw in (12.124), it is necessary also to include spacetime translations. Now, a local spacetime translation $x^\mu \rightarrow x^\mu + a^\mu(x)$, with some arbitrary 4-vector $a^\mu(x)$, amounts to a general coordinate transformation of the kind that we dealt with in chapter 2. Therefore, a theory that is *locally* supersymmetric must also be invariant under general coordinate transformations, so it must include gravity. It is, in short, a *supergravity* theory. One of the gauge fields associated with the combined local symmetry of general coordinate transformations and supersymmetry must be the metric tensor $g_{\mu\nu}(x)$ (or, more or less equivalently, the vierbein (7.130)) and we have seen in §7.6.2 that the corresponding particles are gravitons, which have spin 2. The other is a 'Rarita-Schwinger' field $\psi_\mu(x)$, which is 4-vector, each of whose components is a spinor. A field of this kind describes a spin-$\frac{3}{2}$ particle which, in accordance with our discussion in §12.7.4, is the superpartner of the graviton—the 'gravitino'. The field $\psi_\mu(x)$ has 16 components but, as with the fields for spin-1 and spin-2 particles, not all of these components represent real, independent physical degrees of freedom. In fact, a massive spin-$\frac{3}{2}$ particle has just four helicity states, $h = \pm\frac{1}{2}, \pm\frac{3}{2}$, while a massless one has only the states with $h = \pm\frac{3}{2}$ available to it. In a supergravity theory with unbroken supersymmetry, the gravitino is massless. When local supersymmetry is spontaneously broken, via a 'super-Higgs' mechanism, the two degrees of freedom associated with the Goldstone fermion appear as the extra $h = \pm\frac{1}{2}$ states of a massive gravitino, rather than as an independent massless particle.

I pointed out earlier that although it is quite straightforward to write down a generally-covariant Lagrangian density, which on the face of it describes a quantum theory of gravitational forces, such a theory is not renormalizable. Since then, we have learned that supersymmetry leads to cancellations of potentially divergent terms in perturbation theory. We might wonder, then, whether supergravity, although not renormalizable by the criteria of chapter 9, might actually be *finite*, in the sense that all the potential divergences might cancel. Superficially, we might speculate that the chances of this happening would be improved if our theory had the greatest possible amount of symmetry. There are, as I mentioned in §12.7.4, extended versions of supersymmetry in which N independent supersymmetry transformations are permitted, and correspondingly there are N sets of generators Q_α^a, with $a = 1, \ldots, N$. In such a theory, there are N helicity-raising and helicity-lowering operators of the kind that appeared in (12.135), which means that the helicities of states in an extended-supersymmetry multiplet can vary from $-h_{\max}$ to $+h_{\max}$ in N steps of $\frac{1}{2}$. Simple arithmetic shows, therefore, that $h_{\max} = N/4$. If we allow no fundamental particles with spins greater than the graviton's spin of 2 (and there are theoretical reasons for believing that no fundamental particles with higher spins are possible), then the maximum number of supersymmetries is $N = 8$. In a theory having this maximum degree of symmetry, only one supermultiplet is possible, and it turns out that the Lagrangian of this theory is uniquely determined as well. In the 1980s, considerations of this kind encouraged the hope that $N = 8$ supergravity might be the ‘theory of everything’, determined uniquely by symmetry principles. Painstaking investigations have shown this hope to be unfounded, however. It turns out that, although many cancellations of infinities do occur, they are not sufficient to make the theory finite. Moreover, it does not seem to be possible to recover, by means of spontaneous symmetry breaking, a theory whose lighter particles are the ones actually observed.

In summary, the situation we have arrived at is roughly this. On the one hand, the standard model of strong, electromagnetic and weak interactions, supplemented with the *classical* theory of gravity, is consistent with all experimental data, with the possible exception of neutrino masses. On the other, there seem to be compelling theoretical reasons for doubting that these edifices can really describe the world at its most fundamental level, and the apparent unification of coupling constants at 10^{15}–10^{16} GeV looks like an experimental pointer towards some deeper theory. Grand unified theories, whether or not they incorporate supersymmetry, either fail to reproduce the world as we know it, or do so only at the expense of *ad hoc* manoeuvres that leave the resulting theories hardly more plausible than the standard model itself. For many theorists, hopes of a satisfactorily unified theory of the physical world currently reside in string theory, about which I shall have something to say in chapter 15. At the time of writing, though, it is hard to know whether such hopes may eventually prove well founded.

12.7.6 Some algebraic details

Throughout this section, the γ matrices are those of the Weyl representation given in §7.5. In particular, the charge conjugation matrix C has the properties

$$C^\dagger = C^{\mathrm{T}} = C^{-1} = -C \qquad [C, \gamma^5] = [C, P_{\mathrm{L}}] = [C, P_{\mathrm{R}}] = 0. \tag{12.138}$$

The chirality matrix γ^5 has the properties $\gamma^{5\mathrm{T}} = \gamma^{5\dagger} = \gamma^5$, and this implies that

$$P_{\mathrm{L}}^{\mathrm{T}} = P_{\mathrm{L}}^\dagger = P_{\mathrm{L}} \qquad P_{\mathrm{R}}^{\mathrm{T}} = P_{\mathrm{R}}^\dagger = P_{\mathrm{R}}\,. \tag{12.139}$$

Because γ^5 anticommutes with all the γ^μ, we have

$$P_{\mathrm{L}}\gamma^\mu = \gamma^\mu P_{\mathrm{R}} \qquad \text{and} \qquad P_{\mathrm{R}}\gamma^\mu = \gamma^\mu P_{\mathrm{L}}. \tag{12.140}$$

A Majorana spinor is defined by the property $\psi^{\mathrm{c}} = C\bar{\psi}^{\mathrm{T}} = \psi$, so we can deduce that for a Majorana spinor $\bar{\psi} = \psi^{\mathrm{T}}C$. However, its right- and left-handed components are not themselves Majorana spinors. Their Dirac conjugates are given by

$$\bar{\psi}_{\mathrm{L}} = (P_{\mathrm{L}}\psi)^\dagger\gamma^0 = \psi^\dagger P_{\mathrm{L}}\gamma^0 = \bar{\psi}P_{\mathrm{R}} \tag{12.141}$$

because $\gamma^5\gamma^0 = -\gamma^0\gamma^5$, and similarly $\bar{\psi}_{\mathrm{R}} = \bar{\psi}P_{\mathrm{L}}$. A consequence of this is that, for two Majorana spinors ψ_1 and ψ_2,

$$\psi_{1\mathrm{L}}^{\mathrm{T}}C\psi_{2\mathrm{L}} = \psi_1^{\mathrm{T}}P_{\mathrm{L}}\,C\psi_{2\mathrm{L}} = \psi_1^{\mathrm{T}}CP_{\mathrm{L}}\,\psi_{2\mathrm{L}} = \bar{\psi}_1\,P_{\mathrm{L}}\,\psi_{2\mathrm{L}} = \bar{\psi}_{1\mathrm{R}}\psi_{2\mathrm{L}}. \tag{12.142}$$

If M is a 4×4 matrix, then $\psi_1^{\mathrm{T}} M \psi_2$ is a matrix with a single element, which would ordinarily be equal to its own transpose. However, because ψ_1 and ψ_2 are anticommuting objects, we get

$$\psi_1^{\mathrm{T}} M \psi_2 = -\psi_2^{\mathrm{T}} M^{\mathrm{T}} \psi_1 \tag{12.143}$$

the $-$ sign arising from reversing the order of ψ_1 and ψ_2. However, the complex conjugate of this object must be defined in a way that is consistent with Hermitian conjugation: $(AB)^\dagger = B^\dagger A^\dagger$, regardless of commutation properties. Thus, we have

$$(\psi_1^\dagger M \psi_2)^* = (\psi_1^\dagger M \psi_2)^\dagger = \psi_2^\dagger M^\dagger \psi_1 \, . \tag{12.144}$$

Readers who wish to verify the details of the results given in the preceding subsections should find that a patient application of the algebraic miscellany set out here will meet their purpose. Those who wish to verify the transformation (12.90) of the Wess–Zumino Lagrangian will find it helpful to note that

$$(\eta^{\mu\nu} - \gamma^\nu \gamma^\mu)\partial_\mu \partial_\nu \phi(x) = \tfrac{1}{2}[\gamma^\mu, \gamma^\nu]\partial_\mu \partial_\nu \phi(x) = 0. \tag{12.145}$$

Exercises

12.1. Suppose that the state $|\nu_{\mathrm{e}}\rangle$ containing an electron-type neutrino and the state $|\nu_\tau\rangle$ containing a τ-type neutrino are given by

$$|\nu_{\mathrm{e}}\rangle = \cos\theta_\nu |\nu_1\rangle + \sin\theta_\nu |\nu_2\rangle \qquad |\nu_\tau\rangle = -\sin\theta_\nu |\nu_1\rangle + \cos\theta_\nu |\nu_2\rangle.$$

The particles ν_1 and ν_2 are ‘mass eigenstates’, which means that they can exist as particles with definite masses, m_1 and m_2, and thus with definite energies $E_i = \sqrt{p^2 + m_i^2}$. On the other hand, the neutrinos produced in association with electrons or positrons in nuclear reactions are in the state $|\nu_{\mathrm{e}}\rangle$. The angle θ_ν is a mixing angle analogous to the Cabibbo angle θ_{C} in (12.59). Using the Schrödinger picture of time evolution, show that the state $|\nu(t)\rangle$, which is equal to $|\nu_{\mathrm{e}}\rangle$ at the moment $t = 0$ when a neutrino is produced with a definite 3-momentum of magnitude p, is given at a later time by

$$\begin{aligned} |\nu(t)\rangle = {} & \left(\cos^2\theta_\nu \mathrm{e}^{-\mathrm{i}E_1 t} + \sin^2\theta_\nu \mathrm{e}^{-\mathrm{i}E_2 t}\right)|\nu_{\mathrm{e}}\rangle \\ & + \cos\theta_\nu \sin\theta_\nu \left(\mathrm{e}^{-\mathrm{i}E_2 t} - \mathrm{e}^{-\mathrm{i}E_1 t}\right)|\nu_\tau\rangle. \end{aligned}$$

Consider a neutrino produced in the sun, a distance L from the Earth, with a momentum p that is much greater than m_1 or m_2. We can approximate the energies by $E_i \approx p + m_i^2/2p$ and take the neutrinos to travel with essentially the speed of light. A terrestrial detector is sensitive only to neutrinos of type ν_{e}. Show that the ‘survival probability’ $P_{\nu_{\mathrm{e}}}(L)$, of finding the neutrino in the state $|\nu_{\mathrm{e}}\rangle$ on arrival at the Earth is given approximately by

$$P_{\nu_{\mathrm{e}}}(L) \approx 1 - \sin^2(2\theta_\nu)\sin^2(\Delta m^2 L/4p)$$

where $\Delta m^2 = |m_1^2 - m_2^2|$. If the masses are small, then p is essentially the energy of the detected neutrino. The rate at which solar neutrinos are detected falls short of that expected on the basis of the standard theory of nuclear reactions in the sun by some 50–75%. The actual shortfall varies with energy; the various detectors used are sensitive to different energy ranges, corresponding to neutrinos produced by different reactions. 'Neutrino oscillations' of the kind studied here offer one solution to this *solar neutrino problem*, though other effects must also be taken into account, and a more general mixing of three neutrino species should probably be allowed for. This, together with similar phenomena involving neutrinos produced in nuclear reactors or by cosmic ray interactions in the upper atmosphere, suggests quite strongly that neutrinos have small, but non-zero masses.

12.2. Let $\psi = (\psi_1, \psi_0, \psi_{-1})^{\mathrm{T}}$ be a triplet of scalar fields with weak isospin $t = 1$. Show that the matrices that generate isospin rotations of this triplet are

$$t_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad t_2 = \frac{-\mathrm{i}}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \quad t_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Why do these matrices differ from those shown in equations (B.5) and (B.8) of appendix B?

12.3. Consider an extended version of the GWS model where, in addition to the Higgs field (12.18), there is a triplet Higgs field, such as the ψ of the previous exercise, whose vacuum expectation value is $(0, 0, w)^{\mathrm{T}}$. What weak hypercharge must be assigned to ψ? Show that the value of the parameter $\rho = M_{\mathrm{W}}^2/M_{\mathrm{Z}}^2 \cos^2\theta_{\mathrm{W}}$, which is found experimentally to be very close to 1, is given by

$$\rho = \frac{1 + 2w^2/v^2}{1 + 4w^2/v^2}.$$

Aside from the value of this parameter, why could an electroweak theory involving massive fermions not be constructed using ψ as the only Higgs field?

Chapter 13

Solitons and So On

In both statistical mechanics and quantum field theory, the Euler–Lagrange equations for interacting systems, such as (11.26) in the Ginzburg–Landau treatment of a ferromagnet or (8.43) for a non-Abelian gauge theory, are nonlinear equations governing the behaviour of fields such as the magnetization density $M(\boldsymbol{x})$ or the gauge field $A_\mu(\boldsymbol{x})$. Until now, we have dealt with the nonlinearities perturbatively (apart from our qualitative discussion of QCD in §12.5, where we saw that perturbation theory works only at high energies). That is to say, we have identified constant values of the fields that represent the most stable state of the system by minimizing an appropriate potential or free energy, and treated fluctuations about these constant values as excitations that interact only weakly. In quantum field theory (other than confining theories such as QCD), the quantized 'excitations' of the vacuum state are, of course, the particles observed by experimenters.

Quite often, nonlinear differential equations have solutions that are more complicated than small oscillations about a constant value. They are generically referred to as *solitons* (although some authors reserve this word for solutions having special properties that will not greatly concern us) or, for reasons that will emerge, as *topological defects*. In physical terms, we may interpret these solutions as corresponding to spatially inhomogeneous states of our system, or as representing some kind of particle or extended physical object. In condensed-matter physics, phenomena of this kind are well known; quantum field theory strongly suggests that they ought to occur at a more fundamental level also, but there is no experimental evidence that they actually do. In this chapter, I shall discuss a few examples of solitonic objects and of the theoretical ideas that prove useful in understanding them. For more detailed treatments, readers may like to consult, for example, Coleman (1985), Nakahara (1990), Rajaraman (1987), Tinkham (1996), and Vilenkin and Shellard (2000).

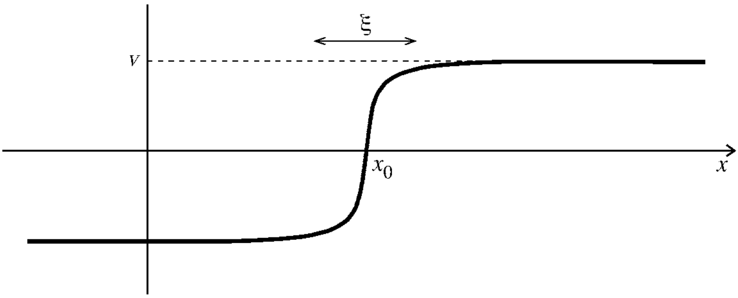


Figure 13.1. Sketch of the domain-wall function (13.3).

13.1 Domain Walls and Kinks

Perhaps the simplest example of a soliton (though it does not meet the more technical definition that I mentioned above) is the domain wall studied in exercise 11.2. In the notation that I shall use throughout this chapter, it is a solution of the equation

$$\nabla^2\phi(\boldsymbol{x}) = -\mu^2\phi(\boldsymbol{x}) + \tfrac{1}{6}\lambda\phi^3(\boldsymbol{x}) = V'(\phi(\boldsymbol{x})) \tag{13.1}$$

where the potential

$$V(\phi) = -\tfrac{1}{2}\mu^2\phi^2 + \tfrac{1}{4!}\lambda\phi^4 \tag{13.2}$$

corresponds to the broken curve in figure 11.6(*b*). If we allow ϕ to vary only in, say, the x direction, then it is easy to verify that

$$\phi_{\mathrm{w}}(x) = v\tanh((x - x_0)/\xi) \tag{13.3}$$

is a solution, where $v = (6\mu^2/\lambda)^{1/2}$ is the positive minimum of the potential and $\xi = \sqrt{2}/\mu$. The function $\phi_{\mathrm{w}}(x)$ is sketched in figure 13.1. As the figure indicates, ξ is a measure of the thickness of the domain wall, which is to say that $\phi(x)$ is almost constant, except in a region whose length is of the order of ξ. This thickness is actually twice the correlation length defined in (11.33).

Beyond the simple fact that it exists, there are several interesting things to be said about this solution. First of all, let us calculate its energy. If we regard ϕ as representing the magnetization of a ferromagnet or the density of a fluid in three dimensions, then the energy density of the system is

$$\mathcal{E}(\phi) = \tfrac{1}{2}\nabla\phi\cdot\nabla\phi - \tfrac{1}{2}\mu^2\phi^2 + \tfrac{1}{4!}\lambda\phi^4 + \tfrac{3}{2}\lambda^{-1}\mu^4 \tag{13.4}$$

where I have added the constant $3\mu^4/2\lambda$ to give the ground states $\phi = \pm v$ an energy density $\mathcal{E}(\pm v) = 0$. From a macroscopic point of view, the plane $x = x_0$ contains a domain wall separating the two coexisting phases, and the sensible thing to calculate is the energy per unit area ϵ of this wall. It is given by

$$\epsilon = \int_{-\infty}^{\infty} \mathrm{d}x\, \mathcal{E}(\phi_{\mathrm{w}}(x)) = \frac{3\mu^4}{\lambda} \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{sech}^4((x - x_0)/\xi) = \frac{4\sqrt{2}\mu^3}{\lambda}. \quad (13.5)$$

We get a finite answer, because $\phi_{\mathrm{w}}(x)$ differs appreciably from $\pm v$, and thus the energy density differs appreciably from zero, only in a region of width ξ. Although this energy per unit area is finite, it is certainly not zero. In fact, the total energy contained in the domain wall is infinite if our system is of infinite size in the y and z directions.

The question arises, is this state of the system a stable one? Generally speaking, we would expect an instantaneous state in which a substantial amount of energy is contained in a restricted region of space to 'dissipate'. With the passage of time, that is to say, a localized 'lump' of energy would be expected to decay, the energy being converted into vibrations of ϕ about the minimum-energy value v. In terms of quantum field theory, a heavy particle generally decays into lighter particles, unless there is some special circumstance that prevents this from happening. It happens that the domain-wall state *is* stable, and it will be instructive to look at this from two points of view: by investigating the effect of small fluctuations about the domain-wall state and by considering its topological properties. The Euler–Lagrange equation (13.1), of which the domain-wall configuration $\phi_{\mathrm{w}}(x)$ is a solution, is the condition for the total energy

$$H_{\mathrm{eff}}(\phi) = \int \mathrm{d}^3x\, \mathcal{E}(\phi) = \int \mathrm{d}^3x \left[\tfrac{1}{2}\nabla\phi \cdot \nabla\phi + \tfrac{1}{4!}\lambda(\phi^2 - v^2)^2\right] \quad (13.6)$$

to have an extremal value. If we think of the solution $\phi_{\mathrm{w}}(x)$ as a point in the infinite-dimensional space of all real functions $\phi(\boldsymbol{x})$, then this point is one at which H_{eff} has a maximum, a minimum or some kind of a saddle point. We see from (13.6) that this energy is a sum of kinetic and potential energies, neither of which can be negative. The absolute minima are the constant solutions $\phi = \pm v$ for which $H_{\mathrm{eff}} = 0$, but $\phi_{\mathrm{w}}(x)$ might be a local minimum. If it is, then any small change in $\phi(\boldsymbol{x})$ will cause the energy to increase. We can check whether this is so by finding the energy of a configuration $\phi(\boldsymbol{x}) = \phi_{\mathrm{w}}(x) + \widetilde{\phi}(\boldsymbol{x})$, assuming that $\widetilde{\phi}(\boldsymbol{x})$ is small. We get

$$\begin{aligned} H_{\mathrm{eff}}(\phi) - H_{\mathrm{eff}}(\phi_{\mathrm{w}}) &= \tfrac{1}{2}\int \mathrm{d}^3x \left[\nabla\widetilde{\phi} \cdot \nabla\widetilde{\phi} - \mu^2\widetilde{\phi}^2 + \tfrac{1}{2}\lambda\phi_{\mathrm{w}}^2\widetilde{\phi}^2\right] + \mathrm{O}(\widetilde{\phi}^3) \\ &= \tfrac{1}{2}\int \mathrm{d}^3x\, \widetilde{\phi}\left[-\nabla^2 - \mu^2 + \tfrac{1}{2}\lambda\phi_{\mathrm{w}}^2\right]\widetilde{\phi} + \mathrm{O}(\widetilde{\phi}^3) \quad (13.7) \end{aligned}$$

where the second version is obtained from an integration by parts. The term linear in $\widetilde{\phi}(x)$ vanishes because we are expanding about an extremum.

A standard method of making sense of this expression requires us to find the eigenfunctions and eigenvalues of the differential operator

$$\mathcal{D} = -\nabla^2 - \mu^2 + \tfrac{1}{2}\lambda\phi_{\mathrm{w}}^2 = -\nabla^2 + 2\mu^2 - 3\mu^2\mathrm{sech}^2((x - x_0)/\xi). \quad (13.8)$$

Since our main interest is clearly in how the state of system varies with position in the x direction, I shall simplify matters from now on by considering the one-dimensional system that we obtain by ignoring the coordinates y and z parallel to the domain wall. With this simplification, we need the functions $f(x)$ and the eigenvalues ω that satisfy the equation

$$\left[-\partial_x^2 + 2\mu^2 - 3\mu^2\mathrm{sech}^2(\bar{x}/\xi)\right] f(x) = \omega^2 f(x) \tag{13.9}$$

where ∂_x means $\partial/\partial x$ and $\bar{x} = x - x_0$. This equation has the same form as the time-independent Schrödinger equation (5.74). It turns out that there are two 'bound states' and a continuous spectrum of 'scattering states'. The eigenfunctions and eigenvalues are

$$f_0(x) = \sqrt{3/4\xi}\,\mathrm{sech}^2(\bar{x}/\xi) \qquad \omega_0^2 = 0 \tag{13.10}$$

$$f_1(x) = \sqrt{3/2\xi}\,\mathrm{sech}(\bar{x}/\xi)\tanh(\bar{x}/\xi) \qquad \omega_1^2 = \tfrac{3}{4}m^2 \tag{13.11}$$

$$f_q(x) = A_q \mathrm{e}^{\mathrm{i}q\bar{x}}\left[3\tanh^2(\bar{x}/\xi) - 1 - q^2\xi^2 - 3\mathrm{i}q\xi\tanh(\bar{x}/\xi)\right] \qquad \omega_q^2 = q^2 + m^2 \tag{13.12}$$

where $m^2 = 2\mu^2 = 4/\xi^2$ and the amplitude A_q is

$$A_q = m^2/4\sqrt{(q^2 + \tfrac{1}{4}m^2)(q^2 + m^2)}. \tag{13.13}$$

These eigenfunctions have the orthonormality properties

$$\int_{-\infty}^{\infty} \mathrm{d}x\, f_0^2(x) = \int_{-\infty}^{\infty} \mathrm{d}x\, f_1^2(x) = 1 \tag{13.14}$$

$$\int_{-\infty}^{\infty} \mathrm{d}x\, f_q(x) f_{q'}(x) = 2\pi\delta(q + q') \tag{13.15}$$

$$\int_{-\infty}^{\infty} \mathrm{d}x\, f_0(x) f_1(x) = \int_{-\infty}^{\infty} \mathrm{d}x\, f_0(x) f_q(x) = \int_{-\infty}^{\infty} \mathrm{d}x\, f_1(x) f_q(x) = 0 \tag{13.16}$$

and it is also true that $f_q^*(x) = f_{-q}(x)$. We now express $\widetilde{\phi}(x)$ as a linear combination of these eigenfunctions

$$\widetilde{\phi}(x) = c_0 f_0(x) + c_1 f_1(x) + \int \frac{\mathrm{d}q}{2\pi} c(q) f_q(x). \tag{13.17}$$

Because $\widetilde{\phi}(x)$ is real, c_0 and c_1 are real, and $c^*(q) = c(-q)$. On substituting this expansion into (13.7), we find

$$H_{\mathrm{eff}}(\phi) - H_{\mathrm{eff}}(\phi_{\mathrm{w}}) = \tfrac{1}{2}\int \mathrm{d}x\, \widetilde{\phi}\mathcal{D}\widetilde{\phi} = \tfrac{1}{2}|c_1|^2\omega_1^2 + \tfrac{1}{2}\int \frac{\mathrm{d}q}{2\pi}\,|c(q)|^2\omega_q^2. \tag{13.18}$$

The fact that all the eigenvalues ω_1^2 and $\omega_{\underline{q}}^2$ are positive tells us that this energy difference is positive. Any small change $\widetilde{\phi}(x)$, specified by the coefficients c_0, c_1 and $c(q)$ leads to an increase in energy, so the domain-wall configuration is indeed a local minimum of the energy.

The reason for the stability of this domain-wall configuration is, in a sense, a topological one. If we insist that the energy (13.5) should be finite (and we normally do insist on this, because a state with infinite energy has a weight of zero in a partition sum such as (11.23)), then $\phi(x)$ can differ significantly from $\pm v$ only over a finite distance. In particular, we must have $\phi(x) \to \pm v$ for $x \to \pm\infty$. The allowed configurations fall into four classes (generally called *sectors*) distinguished by the four possible combinations of boundary conditions. We see that $\phi_{\mathrm{w}}(x)$ has the minimum energy possible in the sector with $\phi(-\infty) = -v$ and $\phi(+\infty) = +v$. Roughly speaking, it achieves this by changing from $-v$ to $+v$ over an optimal distance of the order of ξ. The optimization consists in balancing the cost in potential energy, which increases if the change takes place over a larger distance, against the cost in gradient energy $(\nabla\phi)^2$, which increases if $\phi(x)$ varies more rapidly. The only way to reduce the energy of ϕ_{w} is to change it into a configuration belonging to one of the sectors with $\phi(+\infty) = \phi(-\infty)$. Clearly, this cannot be achieved by adding any small $\widetilde{\phi}(x)$, which accounts for our result that (13.18) is positive. In terms of thermal fluctuations in a system such as a ferromagnet, we can see that a fluctuation which changes the state from one sector to another would have to move an entire half of the system, between x_0 and ∞ across the energy barrier at $\phi = 0$. This requires an infinite energy (or at least a very large energy in a finite but large system) and therefore has an infinitesimal probability of occurring. In principle, the partition function (11.23) is a sum of four parts, say

$$Z = Z_{++} + Z_{+-} + Z_{-+} + Z_{--} \tag{13.19}$$

where Z_{ab} is the integral over configurations for which $\phi(-\infty) = av$ and $\phi(+\infty) = bv$. When we study the statistical mechanics of a classical system, the ensemble average is intended to represent an average over the fluctuations that might occur during the time over which a system is observed, so the relevant partition function is normally the one belonging to just one sector, say Z_{++} for a homogeneous system, or Z_{-+} for a system that contains a domain wall of the kind ϕ_{w}. For a quantum-mechanical system, things may be different, because the state at any instant of time might be a superposition of states belonging to different sectors. Depending on the particular situation we wish to study, the functional integral that represents a probability amplitude might include integrals over several sectors.

The existence of these different sectors of field configurations (or, in quantum mechanics, of vectors in the underlying Hilbert space) is connected, in a way that will become clearer when we look at further examples, with a topological relationship between two spaces. One of these spaces consists of all the points at the spatial boundaries of our system—in this case, the two points

$x = \pm\infty$. The other is the set of values of ϕ at which the potential $V(\phi)$ has its absolute minima. In quantum field theory, these minima correspond to different possibilities for the vacuum state, and this space is called the *vacuum manifold*. In this example, the vacuum manifold also consists of two points, and this fact clearly has a bearing on the nature of the boundary conditions that distinguish the various sectors. In fact, the possibility of having a localized domain wall arises only because of the *im*possibility of changing the value of ϕ continuously from one minimum to another without some large change in potential energy. Consider, indeed, a theory of two fields, ϕ_1 and ϕ_2, with the potential shown in figure 11.8. The field values at $x = \pm\infty$ both lie on the circle of minima, and we can represent the value of ϕ_i at any spatial point x as a point on the potential energy surface. Without calculating an exact solution to the Euler–Lagrange equations, we can see that there is a low-energy configuration $\phi_i(x)$ that interpolates between $\phi_i(-\infty)$ and $\phi_i(+\infty)$ by moving slowly along a path that remains close to the circle of minima as x varies. Its potential energy is always small, and the gradient energy is also small because $\phi_i(x)$ need vary only slowly with x. A domain wall would correspond to a path passing over the central hill, and this is indeed a solution to the Euler–Lagrange equations if the values $\phi_i(\pm\infty)$ are at diametrically opposite points. However, this configuration is now unstable, because the path can be continuously deformed to a low-energy one. In general, a stable domain wall will be possible only if the vacuum manifold has at least two disjoint parts.

It is often convenient to distinguish different sectors according to the value of a *topological charge*. In the present example, it is defined by

$$Q = [\phi(+\infty) - \phi(-\infty)]/2v = (2v)^{-1} \int_{-\infty}^{\infty} \mathrm{d}x\, \partial_x \phi(x). \tag{13.20}$$

A topological charge is always the integral of a total derivative, and thus depends only on the boundary conditions. Here, the topological charges of configurations in the various sectors are $Q_{++} = Q_{--} = 0$, $Q_{-+} = 1$ and $Q_{+-} = -1$.

The eigenfunction $f_0(x)$ given in (13.10) turns out to be especially important. It is no accident that its eigenvalue is exactly zero. In fact, we see that a state of our system containing a domain wall is one in which a continuous symmetry, namely translation invariance, is spontaneously broken. By analogy with the Goldstone bosons that we encountered in §11.7.1, we might expect this state to have a zero-energy (or 'massless') excitation, and clearly it does. A simple proof that this excitation must exist, regardless of the details of the potential $V(\phi)$ that specifies a particular model system, goes like this. The Euler–Lagrange equation (13.1) satisfied by $\phi_{\mathrm{w}}(x)$ is

$$-\partial_x^2 \phi_{\mathrm{w}} + V'(\phi_{\mathrm{w}}) = 0. \tag{13.21}$$

Translation invariance means that $V(\phi_{\mathrm{w}})$ depends on x only through $\phi_{\mathrm{w}}(x)$, so by differentiating this equation, we find

$$\left[-\partial_x^2 + V''(\phi_{\mathrm{w}})\right] \partial_x \phi_{\mathrm{w}} = 0. \tag{13.22}$$

The differential operator $-\partial_x^2 + V''(\phi_{\mathrm{w}})$ is precisely the operator $\mathcal{D}$ in (13.9), so that equation necessarily has a solution $f_0(x)$ proportional to $\partial_x\phi_{\mathrm{w}}(x)$, with eigenvalue equal to zero. Let us make explicit the fact that (13.3) is one of a family of solutions, centred at the point x_0 by writing it as $\phi_{\mathrm{w}}(x - x_0)$. If we change x_0 by a small amount, say δx_0, then a Taylor expansion gives

$$\phi_{\mathrm{w}}(x - x_0 - \delta x_0) \approx \phi_{\mathrm{w}}(x - x_0) - \delta x_0 \partial_x \phi_{\mathrm{w}}(x - x_0) \tag{13.23}$$

so the contribution $c_0 f_0(x)$ to $\widetilde{\phi}(x)$ in (13.17) is a small change in the state of the system that corresponds to moving the position of the domain wall. The function $f_0(x)$ is called the *translation mode*.

Let us now change our point of view, and consider how the soliton might be interpreted in a genuine quantum field theory. If the theory exists in a four-dimensional spacetime, the interpretation is much the same as the one we have already thought about. That is to say, when the symmetry $\phi \rightarrow -\phi$ is spontaneously broken, there may be some regions of the universe in which $\langle 0|\phi|0\rangle = +v$ and others in which $\langle 0|\phi|0\rangle = -v$, and these regions will be separated by domain walls. Within the standard model of particle physics, this interesting possibility does not apply, because the gauge symmetry is a continuous one, with a potential for the Higgs field similar to figure 11.8, and any domain wall would be unstable for the reasons we discussed earlier. Alternatively, we might consider a toy field theory that exists in a spacetime with one spatial dimension. The action for this theory is

$$S = \int \mathrm{d}t\, \mathrm{d}x \left[\tfrac{1}{2}(\partial_t\phi)^2 - \tfrac{1}{2}(\partial_x\phi)^2 - V(\phi)\right] \tag{13.24}$$

and the Euler–Lagrange equation is

$$\partial_t^2\phi - \partial_x^2\phi = -V'(\phi). \tag{13.25}$$

The special feature of this one-dimensional theory is that the soliton now has an energy density that is concentrated near a single point in space, so it might be thought of as some kind of particle. In this context, the soliton solution to the field equation of the $\lambda\phi^4$ theory with potential (13.2) is generally called a *kink*, so I will now denote it by $\phi_{\mathrm{K}}(x)$. The function (13.3) is a solution of the new equation (13.25), which reduces to (13.1) when ϕ is time-independent, and represents a kink that is stationary relative to the (t, x) frame of reference. However, our theory is now Lorentz-invariant, so we ought to be able to find a moving-kink solution by making a Lorentz transformation. In fact, it is simple to verify that

$$\phi_{\mathrm{K}}\left(\frac{x - x_0 - ut}{\sqrt{1 - u^2}}\right) \tag{13.26}$$

is a solution for a kink moving with speed u. The thickness of this moving kink is $\sqrt{1 - u^2}\,\xi$, so the lump of energy that it describes has undergone the Fitzgerald contraction that we might have expected.

Various strategies for treating this and similar models as fully quantum-mechanical systems have been developed in considerable detail (see, for example, Rajaraman (1987) and the original papers cited there). To get everything right is quite a tricky matter, so I shall attempt only to convey some essential ideas. Given a static kink, let us again write $\phi(x,t) = \phi_{\mathrm{K}}(x) + \widetilde{\phi}(x,t)$. Taking account of the equation for ϕ_{K}, the Euler–Lagrange equation becomes

$$\left[\partial_t^2 - \partial_x^2 + V''(\phi_{\mathrm{K}})\right]\widetilde{\phi} = -\tfrac{1}{2}\lambda\phi_{\mathrm{K}}\widetilde{\phi}^2 - \tfrac{1}{6}\lambda\widetilde{\phi}^3. \tag{13.27}$$

The terms on the right-hand side represent interactions between particles and can be ignored to a first approximation if λ is small. The remaining equation is essentially the Klein–Gordon equation (7.2) but with a position-dependent ‘mass’. It is easily solved by using an expansion of the form (13.17) with time-dependent coefficients c_i. The part involving the continuum of eigenvalues ω_q has a straightforward interpretation in terms of the spin-0 particles that would be described by the theory without the kink. For the sake of argument, let us call these particles ‘mesons’ and write

$$\phi_{\mathrm{meson}}(x,t) = \int \frac{\mathrm{d}q}{2\pi}\, c(q,t) f_q(x). \tag{13.28}$$

Since the functions $f_q(x)$ are solutions of (13.9), the equation for $c(q,t)$ is $\partial_t^2 c(q,t) = -\omega_q^2 c(q,t)$ and its solutions are proportional to $\mathrm{e}^{\pm\mathrm{i}\omega_q t}$. We saw earlier that $c^*(q) = c(-q)$, because $\phi(x,t)$ is real, so we can write

$$c(q,t) = (2\omega_q)^{-1}\left[a(q)\mathrm{e}^{-\mathrm{i}\omega_q t} + a^*(-q)\mathrm{e}^{\mathrm{i}\omega_q t}\right] \tag{13.29}$$

and use the fact that $f_q^*(x) = f_{-q}(x)$ to express the meson field as

$$\phi_{\mathrm{meson}}(x,t) = \int \frac{\mathrm{d}q}{2\pi\, 2\omega_q}\left[a(q)\mathrm{e}^{-\mathrm{i}\omega_q t} f_q(x) + a^*(q)\mathrm{e}^{\mathrm{i}\omega_q t} f_q^*(x)\right]. \tag{13.30}$$

This is clearly analogous to the plane-wave expansion (7.11) for a free scalar field and the coefficients $a(q)$ and $a^*(q)$ can be promoted to annihilation and creation operators for mesons in the quantum theory. The functions $f_q(x)$ are, of course, different from the plane-waves $\mathrm{e}^{\mathrm{i}kx}$ for particles of definite momentum k. In fact, they are the wave functions for particles in the potential $U(x) = V''(\phi_{\mathrm{K}}(x))$ provided by the kink. At large distances from the centre of the kink, $x \to \pm\infty$, they reduce to plane waves of the form $\mathrm{e}^{\mathrm{i}qx\mp\mathrm{i}\delta/2}$, where $\delta = 2\tan^{-1}(3q\xi/(2-q^2\xi^2))$, so the mesons do have definite momenta in these distant regions. The angle δ by which the phase of the wavefunction changes as a particle moves through the potential is well known in the quantum theory of scattering by potentials and is called, reasonably enough, the *phase shift*.

At this point, the two degrees of freedom represented by c_0 and c_1 in (13.17) are unaccounted for. The coefficient c_0 of the translation mode is, as it stands, awkward to deal with quantum-mechanically. The reason for this can be appreciated, for example, by interpreting a functional integral such as (9.32) as an integral over the coefficients c_i. The Hamiltonian in (13.7) is independent of c_0, because ω_0 vanishes, and so is the action. Consequently, the integral over c_0 produces a meaningless infinite factor. A means of dealing with this arises from the interpretation of $f_0(x)$ as the first term in the Taylor series (13.23) that shifts the position of the kink. The strategy is to deal with a moving kink, expressing the total field as

$$\phi(x,t) = \phi_{\mathrm{K}}(x - X(t)) + \widetilde{\phi}(x,t) \tag{13.31}$$

where $\widetilde{\phi}(x,t)$ now contains no term proportional to $f_0(x)$. The new degree of freedom $X(t)$ that replaces $c_0(t)$ is called a *collective coordinate*. The action for the kink alone is

$$\begin{aligned} S_{\mathrm{K}} &= \int \mathrm{d}t\, \mathrm{d}x \left[\tfrac{1}{2}(\partial_t\phi_{\mathrm{K}})^2 - \tfrac{1}{2}(\partial_x\phi_{\mathrm{K}})^2 - V(\phi_{\mathrm{K}})\right] \\ &= \int \mathrm{d}t\, \mathrm{d}x \left[\tfrac{1}{2}\dot{X}^2(t)(\partial_x\phi_{\mathrm{K}})^2 - \tfrac{1}{2}(\partial_x\phi_{\mathrm{K}})^2 - V(\phi_{\mathrm{K}})\right]. \end{aligned} \tag{13.32}$$

Because ϕ_K is a function just of $x - X(t)$, the change of integration variable $y = x - X(t)$ eliminates all reference to $X(t)$ except for the factor $\dot{X}^2(t)$ and we find

$$S_{\mathrm{K}} = \int \mathrm{d}t \left[\tfrac{1}{2}M_{\mathrm{K}}\dot{X}^2(t) + \text{constant}\right] \tag{13.33}$$

with the kink mass given by

$$M_{\mathrm{K}} = \xi^{-2}v^2 \int \mathrm{d}y\, \mathrm{sech}^4(y/\xi) = 4\sqrt{2}\,\mu^3/\lambda. \tag{13.34}$$

Not surprisingly, perhaps, this is the same as the energy that we calculated in (13.5). The action S_{K} looks rather like the kinetic energy for a particle of mass M_{K}, and the quantum theory of this model can indeed be interpreted as describing particles of this type, in addition to the mesons. Matters are not entirely straightforward, however. For example, although the form of S_{K} is suggestive, it is actually the action for a *non-relativistic* particle, and this cannot be quite right in a Lorentz-invariant theory. Indeed, the collective coordinate $X(t)$ cannot be interpreted exactly as the position of a moving kink, except perhaps in a non-relativistic limit, because the function $\phi_{\mathrm{K}}(x - X(t))$ does not include the factor $(1 - u^2)^{-1/2}$ that appears in the moving-kink solution (13.26). Moreover, $X(t)$ now appears in place of x_0 in the meson wavefunctions as well as in ϕ_{K} itself, so S_{K} cannot really be considered in isolation. A detailed analysis shows that the momentum conjugate to $X(t)$ is in fact the total momentum for the system, rather than that of the kink on its own, so $X(t)$ itself represents the centre of mass of the entire system, rather than of the kink.

The remaining degree of freedom represented by c_1 can be interpreted in terms of excited states of the kink. In fact, the function $f_1(x)$ in the expansion

(13.17) is the wavefunction for a meson bound in the potential well created by the kink. In contrast to an electron bound in, say, a hydrogen atom, a meson in this theory does not carry any charge that would cancel out the attractive potential of the kink, and since the mesons are bosons, any number of them can occupy the bound state. Consequently, the excited states of the kink are what might be thought of as ‘solitonic atoms’, consisting of the kink with any number of mesons bound to it. If we write $c_1(t)$ (which must be real) as

$$c_1(t) = \frac{1}{\sqrt{2\omega_1}}\left[a_1 \mathrm{e}^{-\mathrm{i}\omega_1 t} + a_1^* \mathrm{e}^{\mathrm{i}\omega_1 t}\right] \tag{13.35}$$

then in the quantum theory the operators a_1 and $a_1^\dagger$ act precisely like the energy lowering and raising operators of the harmonic oscillator; in this case $a_1^\dagger$ adds a bound meson to the atom while a_1 removes one (see exercise 13.2).

13.2 The Sine–Gordon Solitons

It might seem from our discussion in the last section that, even if solitons can be thought of as particles, they must be particles of a quite different kind from those we have dealt with previously. That is, a lump of energy represented by the solution $\phi_{\mathrm{K}}(x)$ seems to be a very different thing from the mesons that are created and annihilated by $a(q)$ and $a^\dagger(q)$. This is not necessarily so, however. Much has been learned from the study of another one-dimensional field theory, the so-called sine–Gordon model, whose Lagrangian density is

$$\mathcal{L} = \tfrac{1}{2}(\partial_t\phi)^2 - \tfrac{1}{2}(\partial_x\phi)^2 + (m^2/\beta^2)\left[\cos(\beta\phi) - 1\right]. \tag{13.36}$$

Its Euler–Lagrange equation is

$$(\partial_t^2 - \partial_x^2)\phi = -(m^2/\beta)\sin(\beta\phi) \qquad \text{or} \qquad (\partial_\mu\partial^\mu + m^2)\phi = \tfrac{1}{6}m^2\beta^2\phi^3 + \ldots \tag{13.37}$$

so m is the mass of the mesons in this theory and β is a coupling constant. The potential $V(\phi) = (m^2/\beta^2)\left[1 - \cos(\beta\phi)\right]$ has an infinity of minima—the candidate vacuum states of the model—at $\phi = 2n\pi/\beta$, for $n = 0, \pm 1, \pm 2, \ldots$ and there are static soliton solutions to the equation of motion (13.37) which interpolate between any neighbouring pair of minima as x varies from $-\infty$ to $+\infty$. Readers should be able to verify, for example, that the function

$$\phi_{1\mathrm{s}}(x) = 4\beta^{-1}\tan^{-1}(\mathrm{e}^{x/\xi}) \tag{13.38}$$

is a solution with the boundary values $\phi(-\infty) = 0$ and $\phi(+\infty) = 2\pi/\beta$, provided that we identify the width parameter as $\xi = 1/m$. As before, the function $\phi_{1\mathrm{s}}\left((x - x_0 - ut)/\sqrt{1-u^2}\right)$ is also a solution, representing a moving soliton whose position is $x_0 + ut$ and whose width is $\bar{\xi} = \sqrt{1-u^2}\,\xi$. The shape of this soliton is qualitatively similar to that of the kink in the $\lambda\phi^4$ theory, though these functions are clearly not exactly the same.

For the sine–Gordon theory, however, many other solitonic solutions can be obtained. Consider, for example, the function

$$\phi_{2\mathrm{s}}(x,t) = \frac{4}{\beta}\tan^{-1}\left[\frac{u\left(\mathrm{e}^{x/\bar{\xi}} - \mathrm{e}^{-x/\bar{\xi}}\right)}{\left(\mathrm{e}^{ut/\bar{\xi}} + \mathrm{e}^{-ut/\bar{\xi}}\right)}\right] \tag{13.39}$$

which has the boundary values $\phi_{2\mathrm{s}}(-\infty,t) = -2\pi/\beta$ and $\phi_{2\mathrm{s}}(+\infty,t) = +2\pi/\beta$. Some straightforward (though somewhat long-winded) algebra will verify that this too is a solution. It describes two moving solitons: one of them interpolates between the minima $n=-1$ and $n=0$, the other between the minima $n=0$ and $n=1$. We can see explicitly how this works by taking the limit $t\to-\infty$, which gives

$$\phi_{2\mathrm{s}}(x,t) \approx 4\beta^{-1}\tan^{-1}\left\{\exp\left[(x+x_0+ut)/\bar{\xi}\,\right] - \exp\left[-(x-x_0-ut)/\bar{\xi}\,\right]\right\} \tag{13.40}$$

where $x_0 = \bar{\xi}\ln u$. In the region of space where x is large and negative, say near x_0+ut, the first exponential is negligibly small, and we have a soliton moving in the positive x direction. Conversely, in the region where x is near $-(x_0+ut)$, the second exponential is negligible and we see a soliton moving in the negative x direction. At very early times, then, we have two widely separated solitons, both moving toward the origin at $x=0$. Similar reasoning shows that at late times, $t\to+\infty$, these two solitons are found moving outwards, having rebounded from each other at a time near $t=0$.

There are, in fact, solutions representing combinations of solitons and antisolitons that interpolate between any two minima of the potential at $x=-\infty$ and $x=+\infty$, so the possible states fall into an infinite number of topological sectors. In this case, the topological charge can be defined as

$$Q = \frac{\beta}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}x\,\partial_x\phi(x,t) \tag{13.41}$$

and it can take any of the values $0, \pm1, \pm2, \ldots$. This charge is independent of t, because it involves only the boundary values $\phi(\pm\infty,t)$ which do not change with time.

Perhaps the most important feature of the sine–Gordon model is the fact, first demonstrated by S Coleman (see Coleman (1985) for an extended discussion and references to the original literature), that it is exactly equivalent, as a quantum field theory, to an apparently quite different model, called the *massive Thirring model*. This is a theory of spin-$\frac{1}{2}$ particles in one space dimension, with the Lagrangian density

$$\mathcal{L}_{\mathrm{T}} = \bar{\psi}(\mathrm{i}\gamma^\mu\partial_\mu - m_{\mathrm{T}})\psi - \tfrac{1}{2}g(\bar{\psi}\gamma^\mu\psi)(\bar{\psi}\gamma_\mu\psi) \tag{13.42}$$

and might be thought of as a toy version of the Fermi theory of §12.1. In a theory with one time and one space dimension, the spinor $\psi(x)$ has just two components and there are two 2×2 γ matrices, which can be taken as $\gamma^0 = \sigma^1$ and $\gamma^1 = -\mathrm{i}\sigma^2$, where σ^i are the Pauli matrices (7.28). Coleman's proof, which is too lengthy for me to reproduce it here, consists in showing that the Green functions of these two theories are identical, provided that the coupling constants are related by

$$\beta^2 = \frac{4\pi}{1 + g/\pi} \tag{13.43}$$

while the masses and field operators are related in such a way that

$$m_{\mathrm{T}}\bar{\psi}\psi \simeq -(m^2/\beta^2)\cos(\beta\phi) \tag{13.44}$$

where $\simeq$ indicates a technicality that I intend to gloss over. The actual relationship between the field operators was worked out by S Mandelstam (1975). It is

$$\psi_1(x,t) \simeq a\exp[-\mathrm{i}\Phi_1(x,t)] \qquad \psi_2(x,t) \simeq -\mathrm{i}a\exp[-\mathrm{i}\Phi_2(x,t)] \tag{13.45}$$

where a is a constant (which includes an infinite renormalization factor as in (9.70)) and the functions $\Phi_i(x,t)$ are

$$\Phi_1(x,t) = \frac{2\pi}{\beta}\int_{-\infty}^{x} \Pi(y,t)\,\mathrm{d}y + \frac{\beta}{2}\phi(x,t) \tag{13.46}$$

$$\Phi_2(x,t) = \frac{2\pi}{\beta}\int_{-\infty}^{x} \Pi(y,t)\,\mathrm{d}y - \frac{\beta}{2}\phi(x,t). \tag{13.47}$$

In these functions, $\Pi(x,t) = \dot{\phi}(x,t)$ is the canonical momentum, which obeys the equal-time commutation relation (7.14). I must emphasize that (13.45) is a quantum-mechanical relation between field operators which do not commute. We cannot recover the Lagrangian density (13.36) of the sine–Gordon theory simply by substituting the fields (13.45) into (13.42). What Mandelstam's somewhat technical analysis does is to show that if $\phi(x,t)$ obeys the equation of motion of the sine–Gordon theory, then $\psi(x,t)$ obeys the equation of motion obtained from (13.42).

The central point of interest is that the 'ordinary' particles created and annihilated by $\psi_i(x,t)$ can be identified with the solitons of the sine–Gordon theory. Remarkably, therefore, although these solitons appear in a bosonic field theory, they are actually fermions. (The same can be shown to be true of the $\lambda\phi^4$ kinks, and exercise 13.3 suggests a simple way of making this plausible, although it does not constitute a proof.) Whether these solitons are spin-$\frac{1}{2}$ particles is a moot point, because angular momentum has no real meaning in a one-dimensional space. I shall now give two straightforward calculations that should serve to indicate how this correspondence works, but I must ask interested readers to consult the literature for more of the technical details. First, let us use the equal-time commutators (7.14) and (7.15) to calculate

$$\begin{aligned} C(x,x') &= [\Phi_1(x,t), \Phi_1(x',t)] \\ &= -\,\mathrm{i}\pi \int_{-\infty}^{x} \delta(y-x')\,\mathrm{d}y + \mathrm{i}\pi \int_{-\infty}^{x'} \delta(y-x)\,\mathrm{d}y \\ &= -\,\mathrm{i}\pi\left[\theta(x-x') - \theta(x'-x)\right] \end{aligned} \tag{13.48}$$

where the step function $\theta(x-x')$ is equal to 0 if $x < x'$ and 1 if $x > x'$. We see that $C(x,x')$ is equal to $-\mathrm{i}\pi$ if $x > x'$ and $+\mathrm{i}\pi$ if $x < x'$. In either case, we have $\exp[C(x,x')] = -1$, and we can apply the result of exercise 5.7(*d*) to show that $\psi_1(x,t)\psi_1(x',t) = -\psi_1(x',t)\psi_1(x,t)$. In fact, similar calculations for the other field components show that

$$\{\psi_i(x,t), \psi_j(x',t)\} = \{\psi_i(x,t), \psi_j^\dagger(x',t)\} = 0 \qquad \text{for } x \neq x' \tag{13.49}$$

so it really is possible to construct anticommuting field operators from commuting ones. Allowing for the possibility that $x = x'$, it is possible to derive the anticommutation relations (7.87), with $\Pi_j = \mathrm{i}\psi_j^\dagger$, but this is rather more difficult because considerable care is needed to deal correctly with products of field operators at the same point. The aim of the second calculation is to find the commutator $[\psi_1(x,t), Q]$, where Q is the topological charge defined in (13.41). As in the derivation of (13.48), we have

$$\left[\Phi_1(x,t), \phi(x',t)\right] = -2\pi\mathrm{i}\beta^{-1}\theta(x-x') \tag{13.50}$$

from which, by the method suggested in exercise 5.3, we can deduce that

$$\left[\psi_1(x,t), \phi(x',t)\right] = -2\pi\beta^{-1}\theta(x-x')\psi_1(x,t). \tag{13.51}$$

Since $\partial_{x'}\theta(x-x') = -\delta(x-x')$, we find

$$[\psi_1(x,t), Q] = \frac{\beta}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}x' \left(\frac{2\pi}{\beta}\right)\delta(x-x')\psi_1(x,t) = \psi_1(x,t). \tag{13.52}$$

This equation is by now very familiar. It has the same form as (5.60) and tells us that $\psi_1(x,t)$ acts on a given state to reduce its topological charge by 1. It does so, according to (13.51), by creating a 'point soliton' of charge -1 at x:

$$\Delta\phi(x') = \frac{2\pi}{\beta}\theta(x-x') = \lim_{\xi\to 0}\frac{4}{\beta}\tan^{-1}(\mathrm{e}^{-(x'-x)/\xi}) \tag{13.53}$$

(see exercise 13.4). Of course, analogous results can be found using any of the operators ψ_i and $\psi_i^\dagger$.

The difference between this point soliton and the solitons of width ξ, which are solutions of the *classical* sine–Gordon equation, can be understood as resulting from a renormalization due to interactions in the quantum theory. Our discussion of QED in §9.7 showed that by keeping only the lowest-order terms in perturbation theory, we get results for quantities such as scattering cross-sections and the Coulomb potential which are essentially the classical ones (except possibly for the effects of the electron's spin, which has no classical analogue). Quantum-mechanical corrections to the classical theory are small if the coupling is weak, as it is in QED, but they will be significant if the coupling is strong. The same applies here. If the coupling constant β of the sine–Gordon theory is small, then the quantum theory can be well described in terms of lumps of energy that are essentially the same as the classical solitons. In this situation, according to (13.43), the coupling constant g of the massive Thirring model is very large. On the other hand, if β^2 is close to 4π (which constitutes a strong coupling in this theory), then the classical solitons cannot be expected to give an accurate picture of the quantum-mechanical excitations. But then g is small, so an alternative picture of almost-free point particles created by ψ becomes quite accurate. The equivalence of these two theories is one of the earliest examples of a phenomenon that has come to be known as *duality*. This term denotes in general the possibility of two apparently different field theories (or, for that matter, statistical-mechanical models) turning out to describe exactly the same physics. Many examples are known, and we shall encounter some of them. The features we have uncovered, that strong coupling in one of the dual theories corresponds to weak coupling in the other, and that solitonic excitations in one theory correspond to point particles of the dual theory seem to be quite characteristic.

13.3 Vortices and Strings

In a system with more spatial dimensions, different possibilities arise. Consider, for example, the statistical-mechanical model whose Hamiltonian is (11.55) with $r_0 < 0$. For the moment, we shall take the number of dimensions to be $d = 2$, so in this chapter's notation it is

$$H_{\mathrm{eff}}(\phi) = \int \mathrm{d}^2x \left[\nabla\phi^* \cdot \nabla\phi + \tfrac{1}{4}\lambda(\phi^*\phi - v^2)^2\right]. \tag{13.54}$$

If we write $\phi(\boldsymbol{x}) = \phi_1(\boldsymbol{x}) + \mathrm{i}\phi_2(\boldsymbol{x})$, then this is equivalent to

$$H_{\mathrm{eff}}(\phi) = \int \mathrm{d}^2x \left[|\nabla\phi_1|^2 + |\nabla\phi_2|^2 + \tfrac{1}{4}\lambda(\phi_1^2 + \phi_2^2 - v^2)^2\right] \tag{13.55}$$

so ϕ can be regarded as equivalent to a vector $\boldsymbol{\phi} = (\phi_1, \phi_2)$. This vector might, perhaps, be interpreted as the magnetization density of a two-dimensional magnetic system, in which case its direction is a direction in space. Alternatively, the complex field $\phi(\boldsymbol{x})$ might be the condensate wavefunction of a superfluid or superconductor, in which case ϕ_1 and ϕ_2 are the components of a vector in a two-dimensional 'internal' space.

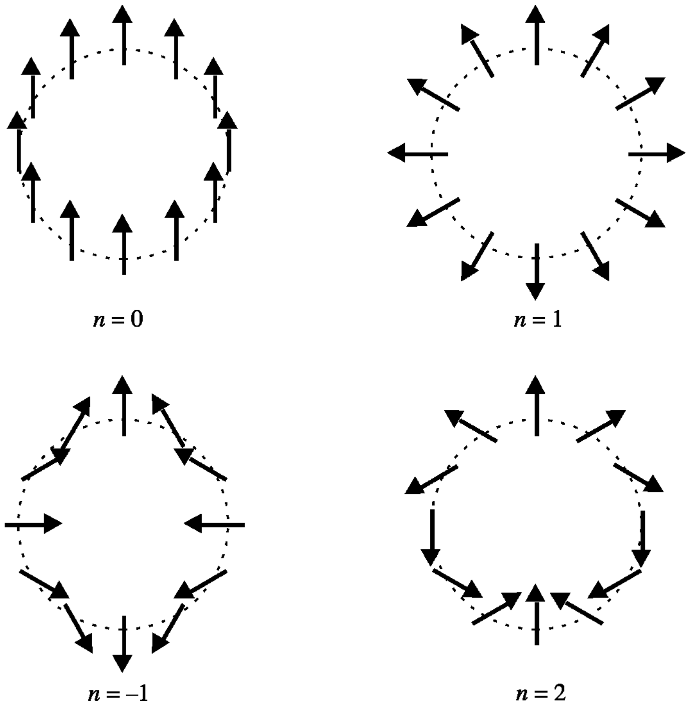


Figure 13.2. Configurations of winding number 0, 1, -1 and 2 for a 2-component field in two spatial dimensions. Depending on the physical interpretation of the field, the arrows might represent the directions in real space of atomic spins or the directions in an internal space which represent the phase angle $\alpha(\boldsymbol{x})$. There should be an arrow at each point in space, but only those at selected points on the dashed circle are drawn.

The vacuum manifold of this model is the circle of minima in figure 11.8. As we discussed earlier, a domain wall described by (13.3) would not be stable, but for a two-dimensional system the boundary at infinity, which we can think of as the limit of a large circle, has the same topology as the vacuum manifold. In any state of finite energy, the magnitude $|\phi(\boldsymbol{x})|$ must, as before, approach the value v as $|\boldsymbol{x}|$ approaches infinity in any direction. Let us again think of the value of $\phi(\boldsymbol{x})$ at a point $\boldsymbol{x}$ on a large circle in space as being represented by a point on the circle of minima. If we move the point $\boldsymbol{x}$ once around its circle in space, the point representing $\phi(\boldsymbol{x})$ must return exactly to its starting point, because ϕ has a unique value at each point. Although the motion of ϕ need not always be in the same direction, overall it must complete a whole number n of circuits. This number is clearly the analogue of the topological charge (13.20); it is often referred to as the *winding number*. Using polar coordinates $\boldsymbol{x} = (r, \theta)$ and a polar representation of the field, $\phi(\boldsymbol{x}) = \rho(\boldsymbol{x}) \exp[\mathrm{i}\alpha(\boldsymbol{x})]$, we can say that on the circle at infinity α is just a function of θ and

$$n = \frac{1}{2\pi}\int_0^{2\pi} \frac{\mathrm{d}\alpha(\theta)}{\mathrm{d}\theta}\,\mathrm{d}\theta. \tag{13.56}$$

Alternatively, if we think of $\alpha(\boldsymbol{x})$ as representing the direction of an atomic spin in a magnet, then the spins that live on a large circle can be pictured as the arrows in figure 13.2, which illustrates states with winding numbers $n = 0, \pm 1$ and 2. Imagine now what happens if we look at the field $\phi(\boldsymbol{x})$ on smaller and smaller circles. It varies continuously with position, and if the winding number is non-zero, then a little thought will show that there must be at least one point in space, say $\boldsymbol{x}_0$, at which α has all its values at once or, in other words, is not well defined. This is possible only if $\rho(\boldsymbol{x}_0) = 0$, which corresponds to the maximum of the potential. Thus, there is a lump of energy centred at $\boldsymbol{x}_0$, which is called a *vortex*. The simplest state, containing a single vortex at the origin (and, as we might guess, the state of lowest energy for winding number n), has the radially symmetric form

$$\phi_{\mathrm{v}}(r,\theta) = \rho(r)\mathrm{e}^{ni\theta}. \tag{13.57}$$

Given a state of this kind, the Euler–Lagrange equation for $\rho(r)$ is

$$\frac{\mathrm{d}^2\rho}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}\rho}{\mathrm{d}r} - \frac{n^2}{r^2}\rho = -\frac{m^2}{2}\rho + \frac{\lambda}{2}\rho^3 \tag{13.58}$$

where $m = \lambda^{1/2}v$ is the mass of the excitations we met in §11.7.1 that correspond to fluctuations in the magnitude of ϕ. In contrast to the one-dimensional equations of previous sections, no exact solution to this equation is known. However, a solution with the right boundary conditions, $\rho \to v$ as $r \to \infty$ and $\rho \to 0$ as $r \to 0$ can be obtained numerically. The energy of this vortex,

$$H_{\mathrm{eff}}(\phi_{\mathrm{v}}) = 2\pi\int_0^\infty \mathrm{d}r\,r\left[(\partial_r\rho)^2 + n^2r^{-2}\rho^2 + \tfrac{1}{4}\lambda\left(\rho^2 - v^2\right)^2\right] \tag{13.59}$$

is infinite. The first and third terms give finite integrals, because (see exercise 13.5) the large-r behaviour of ρ is $\rho(r) = v + \mathrm{O}(r^{-2})$. In the second term, suppose that ρ differs only negligibly from v if $r > a$. The exact value of a for which we might be satisfied with this approximation does not really matter. The contribution to the energy that we get by integrating this term from a to some large distance R is

$$E(R) \approx 2\pi n^2v^2\int_a^R \mathrm{d}r\,r^{-1} \approx 2\pi n^2v^2\ln(R/a) \tag{13.60}$$

and we see that this becomes infinite when $R \to \infty$.

A single vortex, therefore, is not an allowed solution to the Euler–Lagrange equation which, let us remember, is the condition for the energy (13.54) to be a minimum. Consider, however, a configuration of $\phi(\boldsymbol{x})$ that contains a vortex of charge $n = 1$, centred at $\boldsymbol{x}_1$, and another of charge $n = -1$ centred at $\boldsymbol{x}_2$. We might call this a vortex-antivortex pair. In a little more detail, this means the following. Let us define a winding number associated with a closed curve C by

$$n_C = (2\pi)^{-1} \oint_C \nabla\alpha(\boldsymbol{x}) \cdot \mathrm{d}\boldsymbol{\ell} \tag{13.61}$$

where $\mathrm{d}\boldsymbol{\ell}$ is an infinitesimal tangent vector to C. This measures the total change in the phase angle $\alpha(\boldsymbol{x})$ when $\boldsymbol{x}$ is taken once around the curve, and it reduces to (13.56) if C happens to be a circle centred on the origin. In the state we are thinking of, n_C will be equal to 1 if C encloses $\boldsymbol{x}_1$ but not $\boldsymbol{x}_2$, equal to -1 if C encloses $\boldsymbol{x}_2$ but not $\boldsymbol{x}_1$, and equal to 0 if C encloses both or neither of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The winding number measured on the circle at infinity is zero, so the total energy of this state is finite, provided that $\rho(\boldsymbol{x})$ approaches v fast enough as $|\boldsymbol{x}| \to \infty$. There will be many states in which the winding number has these properties. In general, none of these states is a solution of (13.58), but we can we can think of an idealized state containing a vortex, an antivortex and nothing else as one that minimizes the energy (13.54) subject to the constraint that the winding numbers are those we have specified, given two fixed points which are the centres of the vortices. It is, at least, a state of finite energy, because the winding number vanishes on the circle at infinity. Suppose, though, that the vortex and antivortex are separated by a large distance. Near the vortex at $\boldsymbol{x}_1$, the function $\phi(\boldsymbol{x})$ will be almost the same as if the antivortex did not exist. The energy contained in a circle of radius R centred on $\boldsymbol{x}_1$ will increase roughly as $\ln(R/a)$ as R increases, until R becomes comparable with the separation $|\boldsymbol{x}_1 - \boldsymbol{x}_2|$. These qualitative considerations should make it plausible that the energy of the static vortex-antivortex state increases roughly as $\ln|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ as the separation is increased. This can be confirmed by more detailed calculations. In fact, we might add more vortex-antivortex pairs, and the vortices behave in much the same way as particles with a potential energy $V(r) \propto \ln(r/a)$ between particles separated by a distance r. This is, in fact, equivalent to a *Coulomb gas* of electrically charged particles, because in two dimensions the solution of Poisson's equation $\nabla^2 V(\boldsymbol{x}) = q\delta(\boldsymbol{x})$ for the potential due to a point charge at the origin is $V(r) = (q/2\pi)\ln(r/a)$ (see exercise 13.6). Because of this potential energy, there is a force acting between any pair of vortices, so we would not expect to find genuine static solutions to the equations of motion.

It is worth mentioning that these vortices have played a pivotal role in the understanding of a class of phase transitions in two-dimensional systems. A celebrated theorem due to N D Mermin and H Wagner (1966) asserts that continuous symmetries cannot be spontaneously broken in a two-dimensional system (or, for that matter, in a one-dimensional system either). Roughly speaking, the reason is that Goldstone-mode fluctuations in two dimensions (or in one dimension) are sufficiently strong that all the minima of a potential such as the one in figure 11.8 contribute equally to the statistical sum that must be carried out to determine the expectation value $\langle\boldsymbol{\phi}(\boldsymbol{x})\rangle$, with the result that this expectation value vanishes. In higher dimensions, these fluctuations are also present. They result in corrections to the value of $\langle \boldsymbol{\phi}(\boldsymbol{x}) \rangle$ obtained by minimizing $H_{\text{eff}}(\boldsymbol{\phi})$, which can be estimated by the renormalization-group methods of §11.6, but do not destroy the ordered state completely. The model (13.54) that we are currently studying is a version of what is known in statistical mechanics as the *two-dimensional XY model*. Because of its special topological properties, it is found to have a phase transition, even though $\boldsymbol{\phi}$ does not acquire a non-zero expectation value. According to an analysis first given by J M Kosterlitz and D J Thouless, the 'ordered' phase is one in which vortex-antivortex pairs are tightly bound, while in the 'disordered' phase vortices and antivortices can move at random through the system. It turns out that a surprisingly large class of two-dimensional model systems are more or less equivalent, at least as far as their phase transitions are concerned. Our discussion above indicates in outline a correspondence between the XY model and a gas of charged particles. These models are also related to the two-dimensional sine–Gordon theory which, as we discovered earlier, is itself equivalent to a theory of fermions. (In this context, we need the Euclidean version of the sine–Gordon theory, interpreting the time coordinate of §13.2 as $t = \mathrm{i}x^2$, where x^2 is a second spatial coordinate.) One way of seeing this relationship is to consider (13.54) as a model for a magnetic system. By taking λ to be very large, we make the minima in the potential very deep, so that $|\phi|$ is essentially constrained to be equal to the constant v, which we can take to be $v = 1$. In that case, we have $\nabla\phi^* \cdot \nabla\phi \approx |\nabla\alpha|^2$. A small magnetic field in the direction of ϕ_1 gives rise to a potential energy $\boldsymbol{h} \cdot \boldsymbol{\phi} = h \cos\alpha$ and these two terms yield an effective Hamiltonian of the sine–Gordon form. On the other hand, it can be shown directly that the sine–Gordon model is equivalent to a grand-canonical description of a gas of charged particles, the coefficient of $\cos(\beta\phi)$ being equal to the fugacity (10.24). Readers who wish to pursue the details of these matters will find a large and interesting literature waiting to be explored. The paper by Samuel (1978) and the review article by Nienhuis (1987) may provide a useful starting point.

The considerations of this section so far can be generalized in two important ways. First, it is a simple matter in principle to add more dimensions. In a three-dimensional space, say with coordinates (x, y, z), we can consider field configurations for which ϕ is independent of z. A vortex centred at $x = y = 0$ is now a lump of energy that occupies the entire z axis; it is a *vortex line* or a *string*. Its topology is characterized by the winding number (13.61), which has the same value for any curve C that encircles the string exactly once. The vortex energy (13.59) is now the energy per unit length of the string (or, in a relativistic theory, the mass per unit length), which is often referred to as the *string tension*. (The energy stored in an ordinary elastic string when it is stretched is $E = \int T\mathrm{d}x$, where T is the tension. This tension generally increases as the string is stretched, in an ideal case according to Hooke's law, but the tension of the strings we are thinking about is independent of their length.) Again, the energy per unit length of a single straight string is infinite. One can, however, envisage a three-dimensional network of strings that has a finite total energy. The strings in such a network need not be straight; they may indeed form closed loops. As with the two-dimensional vortices, a network of strings does not constitute a static, energy-minimizing state. If such a network is formed by some non-equilibrium process, it will evolve with time in a manner that is not easy to determine in general.

To see how a network of strings might come into existence, consider the fact that cooling a ferromagnetic material from a temperature above its critical temperature T_c to one below T_c typically results in a state containing many domains, in which the magnetization points in different directions. In outline, the reason is that the instantaneous directions of atomic spins at the instant that the temperature passes through T_c are well correlated only over distances smaller than a correlation length ξ_{ne}. This non-equilibrium correlation length is not very precisely defined; it depends on factors such as the rate of cooling in ways that are hard to discover with any degree of rigour. However, it is not the same as the equilibrium correlation length (11.16), which becomes infinite at T_c, unless the cooling process is extremely slow. Below the transition temperature, therefore, the magnetization density will, at least initially, be uniform only over distances of the order of ξ_{ne} and we may expect to find domains of roughly this size, separated by domain walls. A system whose vacuum manifold is the circle of minima in figure 11.8 will not form domains, for the reasons we have discussed. Consider, however, the instantaneous state of such a system as its temperature passes through T_c and, in particular, an arbitrary closed curve C whose length is considerably greater than ξ_{ne}. There is a good probability (though again one that is difficult to quantify precisely) that the winding number on this curve will be non-zero, on account of the random variations in $\boldsymbol{\phi}(\boldsymbol{x})$ over distances greater than ξ_{ne}. After further cooling, $\boldsymbol{\phi}(\boldsymbol{x})$ is increasingly constrained to have values near the vacuum manifold. Thermal fluctuations have too little energy for $\boldsymbol{\phi}(\boldsymbol{x})$ to surmount the energy barrier over large regions, so in the short term the topology of the field configuration is 'frozen in' and a curve which has a non-zero winding number must be found to encircle at least one string. In condensed matter physics, this picture is borne out by experimental observations of temporary string networks formed by rapid 'quenching' of liquid helium through its superfluid phase transition. The possibility of a similar phenomenon in relativistic field theories was emphasized in an influential paper of T W B Kibble (1976). In this case, it seems likely that the requisite phase transitions may have occurred in the very early universe, and I shall return briefly to this in the next chapter. In the cosmological context, the formation and evolution of string networks have been investigated quite extensively by approximate methods, as is discussed in detail by Vilenkin and Shellard (2000).

The second generalization is to consider gauge-invariant theories, of which an extremely important example is the Ginzburg–Landau superconductor. In our present notation, the effective Hamiltonian (11.59) for the superconductor is

$$H_{\text{eff}} = \int \mathrm{d}^3x \left[\tfrac{1}{2}B^2 + |(\nabla - \mathrm{i}q\boldsymbol{A})\phi|^2 + \tfrac{1}{4}\lambda(\phi^*\phi - v^2)^2 - \boldsymbol{B}\cdot\boldsymbol{H} \right] \qquad (13.62)$$

with $q = 2e$ and $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$, and we have a pair of Euler–Lagrange equations

$$(\boldsymbol{\nabla} - \mathrm{i}q\boldsymbol{A})^2\phi = \tfrac{1}{2}\lambda(\phi^*\phi - v^2)\phi \tag{13.63}$$

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) = -\mathrm{i}q\left(\phi^*\boldsymbol{\nabla}\phi - \phi\boldsymbol{\nabla}\phi^* - 2\mathrm{i}q\boldsymbol{A}\phi^*\phi\right) \tag{13.64}$$

which in this context are called the *Ginzburg–Landau equations*. It was discovered by A A Abrikosov as long ago as 1957 that these equations have solutions corresponding to vortex lines, in whose cores there is a magnetic flux. In the context of relativistic field theories, these vortices are commonly associated with the names of H B Nielsen and P Olesen, who rediscovered them somewhat later. Let us again look for a radially symmetric solution of the form (13.57), with $r = \sqrt{x^2 + y^2}$ and take the vector potential to have the form

$$\boldsymbol{A}(\boldsymbol{x}) = nr^{-2}f(r)\,(-y, x, 0) \tag{13.65}$$

so that both ϕ and $\boldsymbol{A}$ are independent of z. The magnetic induction is in the z direction and its strength is

$$B = (\boldsymbol{\nabla} \times \boldsymbol{A})_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} = \frac{n}{r}\frac{\mathrm{d}f}{\mathrm{d}r}. \tag{13.66}$$

With these assumptions, the Ginzburg–Landau equations are

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r} - \frac{n^2}{r^2}\left(1 - qf(r)\right)^2\right]\rho(r) = \frac{\lambda}{2}\left[\rho^2(r) - v^2\right]\rho(r) \tag{13.67}$$

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{r}\frac{\mathrm{d}f(r)}{\mathrm{d}r}\right) = -\frac{2q}{r}\left(1 - qf(r)\right)\rho^2(r). \tag{13.68}$$

Again, no exact solution to these equations can be written down. For small values of r, we can solve for $\rho(r)$ and $f(r)$ as power series in r, and it is not hard to find the limiting behaviour

$$\rho(r) \approx \rho_0 r^{|n|} \qquad f(r) \approx f_0 r^2 \qquad B(r) \approx 2nf_0. \tag{13.69}$$

For $r \to \infty$, on the other hand, one finds

$$\rho(r) \approx v - \rho_\infty \mathrm{e}^{-mr} \quad f(r) \approx q^{-1} - f_\infty r^{1/2}\mathrm{e}^{-\mu r} \quad B(r) \approx n\mu f_\infty r^{-1/2}\mathrm{e}^{-\mu r} \tag{13.70}$$

with $m = \lambda^{1/2}v$ and $\mu = \sqrt{2}qv$. The constants of integration ρ_0, f_0, ρ_∞ and f_∞ cannot be determined analytically, but numerical solutions with this limiting behaviour can be obtained. In contrast to our previous vortex solution, we see that $\rho(r)$ and $B(r)$ approach their large-distance values v and 0 exponentially fast. The distances that characterize the exponential decay are the coherence length $\xi = 1/m$ and the penetration depth $\lambda_\mathrm{p} = 1/\mu$ that we encountered in §11.7.3. In a relativistic field theory, of course, m and μ are respectively the masses of a

Higgs particle and a gauge boson. Because of this exponential decay, the energy per unit length of a vortex line is now finite. It is given by

$$E = 2\pi \int_0^\infty \mathrm{d}r\, r \left[\tfrac{1}{2}B^2 + \left(\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)^2 + (1 - qf)^2 \frac{n^2\rho^2}{r^2} + \tfrac{1}{4}\lambda\left(\rho^2 - v^2\right)^2 \right] - H\Phi \tag{13.71}$$

where $\Phi = \int \mathrm{d}^2r\, B(r)$ is the total magnetic flux passing through the vortex. This magnetic flux can be found exactly, even though we have no exact expression for $B(r)$. To calculate it, we use Stokes' theorem to write

$$\Phi = \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \int_S (\nabla \times \boldsymbol{A}) \cdot \mathrm{d}\boldsymbol{S} = \int_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{\ell} \tag{13.72}$$

where C is a very large circle in the $x-y$ plane and S is the disc that it encloses. On this circle, we have $f(r) = q^{-1}$ and from (13.65) the vector potential is $\boldsymbol{A}(\boldsymbol{x}) = (n/qr^2)(-y, x, 0)$. It is a simple exercise using polar coordinates to calculate

$$\Phi = \int_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{\ell} = 2\pi n/q \tag{13.73}$$

and we see that the flux is n times a universal *flux quantum* $\Phi_0 = 2\pi/q$, whose value is independent of the constants λ and v that characterize a particular superconducting material. In SI units, the flux quantum is $\Phi_0 = h/2e = 2.07 \times 10^{-15}$ Wb. In fact, this flux quantum appears under more general circumstances than the vortex state we are thinking of here. Let C be any closed curve in a region of a superconductor where $\phi^*\phi = v^2$ and $B = 0$. We can write $\phi(\boldsymbol{x}) = v \exp[\mathrm{i}\alpha(\boldsymbol{x})]$ and the second Ginzburg–Landau equation (13.64) becomes $\boldsymbol{A}(\boldsymbol{x}) = q^{-1}\nabla\alpha(\boldsymbol{x})$. Thus, the flux passing through C is just

$$\Phi = \int_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{\ell} = q^{-1} \int_C \nabla\alpha(\boldsymbol{x}) \cdot \mathrm{d}\boldsymbol{\ell} = 2\pi n_C/q \tag{13.74}$$

where n_C is the winding number (13.61). If, for example, we have a sample of superconducting material with a hole in it, then the total flux passing through the hole is quantized in units of Φ_0 and this flux quantization arises simply from the fact that the condensate wavefunction $\phi(\boldsymbol{x})$ must be single-valued.

Under appropriate circumstances, vortex lines are indeed observed in real superconductors. In an ideal case, what is actually seen is a *flux lattice*, as originally predicted by Abrikosov, who found approximate solutions to the Ginzburg–Landau equations in which vortices form square or triangular arrays. To determine which of these arrays is the more stable requires careful numerical calculations of their energies, and the triangular arrays that are actually observed do turn out to be marginally the more stable (though Abrikosov himself originally came to the opposite conclusion). A superconductor containing a stable flux lattice is said to be in a *mixed state*, and the question arises whether this mixed state is more stable than either of the homogeneous states we met in §11.7.

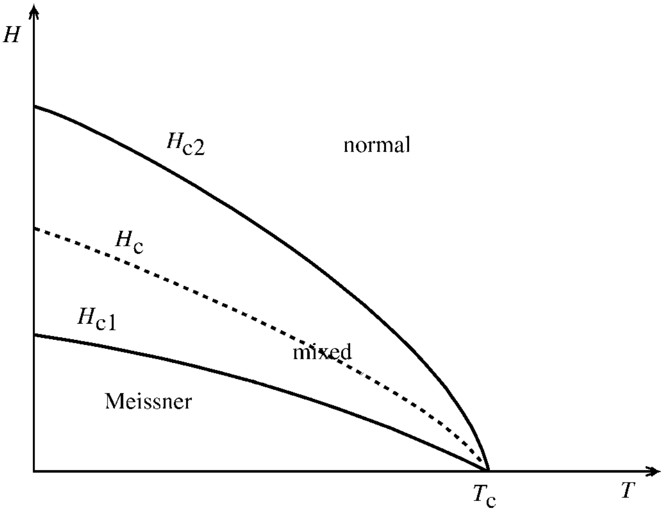


Figure 13.3. Schematic phase diagram of a type-II superconductor.

There, we found that the normal state, with $\phi = 0$ and $B = H$ is the more stable if the externally applied magnetic field H is greater than the critical value $H_c = (\lambda v^4/2)^{1/2}$ while the superconducting state with $|\phi| = v$ and $B = 0$ (which I shall call the 'Meissner state') is the more stable if $H < H_c$. In a rough and ready way, we can think of a vortex as having a core, in which the material is in its normal state, separated from the surrounding superconducting region by a cylindrical wall. Suppose that $H < H_c$. The presence of normal-state regions in the cores of vortices tends to increase the energy (or, more accurately, the free energy) but this might be offset if the free energy of the walls were negative. It turns out (see Tinkham (1996) for details) that this is so if $\lambda_p > \xi$. In that case, the mixed state is more stable than the Meissner state when $H_c > H > H_{c1}$, where H_{c1} is a lower critical field determined essentially by the fact that at least one flux quantum Φ_0 must be available to form each vortex. If $H > H_c$, then the mixed state may be more stable than the normal state, because now the excess free energy of the superconducting regions is offset by the negative energy of the vortex walls. In fact, this happens for $H_c < H < H_{c2}$, where H_{c2} is an upper critical field, whose significance is roughly this. As H increases, so does the total flux penetrating the superconductor, so the vortices become more densely packed. At $H = H_{c2}$, they merge completely, and the mixed state becomes indistinguishable from the normal state. It is conventional to classify superconducting materials according to the *Ginzburg–Landau parameter*

$$\kappa = \lambda_p/\xi = m/\mu = (\lambda/2q^2)^{1/2}. \tag{13.75}$$

(In much of the literature on superconductivity, however, definitions of ξ and hence of κ are used which differ from mine by a factor of $\sqrt{2}$.) If $\kappa < 1$, then the energy of the vortex walls is positive and the mixed state is never stable. A material of this kind is called a *type-I* superconductor. A *type-II* superconductor has $\kappa > 1$ and exhibits the mixed state when $H_{c1} < H < H_{c2}$. When κ is large, the upper and lower critical fields are found to be given roughly by

$$H_{c1} \approx \kappa^{-1} H_c \qquad H_{c2} \approx \kappa H_c. \tag{13.76}$$

The value of H_c depends on temperature, because v does. The details of this temperature dependence cannot, as we saw in §11.7, be determined from the Ginzburg–Landau theory itself, but it must vary as $H_c \sim (T_c - T)$ near the critical temperature. From the considerations above, we see that the phase diagram of a type-II superconductor is that shown schematically in figure 13.3.

We might expect that strings analogous to the vortices observed in superconductors should exist in the non-Abelian gauge theories of particle physics. Typically, the vacuum manifolds of these theories are more complicated than the circle of minima in figure 11.8, so the topological criterion for the existence of strings must be formulated in a more general way. Consider once more a closed curve C in space which, we hope, encircles a string. As $\boldsymbol{x}$ moves once around C, the point $\phi(\boldsymbol{x})$ traces out a closed curve on the vacuum manifold, say C_{vm}. To determine whether C does encircle a topologically stable string, we attempt to shrink C to a point, and this will entail shrinking C_{vm} to a point also. As we do this, we allow the function ϕ to change continuously, so as to ensure that $\phi(\boldsymbol{x})$ is indeed a point on the vacuum manifold whenever $\boldsymbol{x}$ is a point on C. If we can do this successfully, starting with *any* closed curve C, it follows that ϕ can be changed continuously until its value at every point in space is on the vacuum manifold, so there is no stable string. The only circumstance that might prevent us from shrinking C to a point is that C_{vm} cannot be continuously shrunk to a point on the vacuum manifold. Thus, topologically stable strings are possible if there is some closed curve on the vacuum manifold that cannot be shrunk continuously to a point. The mathematical jargon for this circumstance is that the *first homotopy group* of the vacuum manifold is non-trivial. If the vacuum manifold is a circle, then no curve C_{vm} that travels at least once around it can be shrunk to a point, so the strings we have considered do satisfy this criterion. In the standard electroweak theory, the Higgs field (12.18) has four real components, say $\phi_1, \ldots, \phi_4$ and the vacuum manifold is the 3-dimensional surface defined by

$$\phi^\dagger \phi = \phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2 = v^2. \tag{13.77}$$

It is the 3-dimensional surface S^3 of a sphere in four dimensions. Any closed curve on this spherical surface can be shrunk continuously to a point, so strings are not possible in this theory. They are possible in a variety of grand-unified theories, however, and in this context are generally called *cosmic strings*, on the grounds that they may have been produced in the early universe and that very long strings might still exist in the present universe. The characteristic energy scale of symmetry breaking in grand unified theories is $M_{\mathrm{X}} \sim 10^{16}\,\mathrm{GeV}$ and the associated characteristic length scale in natural units is M_{X}^{-1}, so dimensional analysis suggests a mass per unit length for these strings of the order of $\epsilon \sim M_{\mathrm{X}}^2$ or in SI units $\epsilon \sim M_{\mathrm{X}}^2 c/\hbar \sim 10^{21}\mathrm{kg\,m^{-1}}$. Clearly, these objects are very heavy. They might perhaps be detected through a 'gravitational lensing' effect analogous to the bending of light by the sun (see exercise 4.4) but none have been seen up to now.

13.4 Magnetic Monopoles

We saw in §3.7 that Maxwell's equations can be expressed as $\mathrm{d}\boldsymbol{F} = 0$ and $\mathrm{d}\,{}^*\boldsymbol{F} = {}^*\boldsymbol{j}$, where ${}^*\boldsymbol{F}$ and ${}^*\boldsymbol{j}$ are the dual tensors to the field strength tensor $\boldsymbol{F}$ and the electric current $\boldsymbol{j}$. The term 'dual' here refers to the tensor operation specified by (3.82) and (3.83), but we can see that a duality of the kind that we met in connection with the sine–Gordon and massive Thirring models is also involved. In what is perhaps less esoteric language, the components ${}^*F_{\mu\nu}$ given in (3.91) are obtained from those of $F_{\mu\nu}$ by making the replacements $\boldsymbol{B} \to \boldsymbol{E}$ and $\boldsymbol{E} \to -\boldsymbol{B}$. Let us, then, define

$$\widetilde{\boldsymbol{E}} = \boldsymbol{B} \qquad \widetilde{\boldsymbol{B}} = -\boldsymbol{E} \qquad \widetilde{\rho}_{\mathrm{m}} = -\rho_{\mathrm{e}} \qquad \widetilde{\boldsymbol{j}}_{\mathrm{m}} = -\boldsymbol{j}_{\mathrm{e}}. \tag{13.78}$$

With this notation (and with $c = 1$), Maxwell's equations (3.44)–(3.47) become

$$\begin{gathered}\nabla \cdot \widetilde{\boldsymbol{E}} = 0 \qquad \nabla \cdot \widetilde{\boldsymbol{B}} = \widetilde{\rho}_{\mathrm{m}} \\ \nabla \times \widetilde{\boldsymbol{E}} + \frac{\partial \widetilde{\boldsymbol{B}}}{\partial t} = -\widetilde{\boldsymbol{j}}_{\mathrm{m}} \qquad \nabla \times \widetilde{\boldsymbol{B}} - \frac{\partial \widetilde{\boldsymbol{E}}}{\partial t} = 0.\end{gathered} \tag{13.79}$$

Evidently, our conventional theory of electromagnetism, in which there are electrically charged particles but no magnetic monopoles is equivalent to a dual theory in which there are magnetic monopoles, with magnetic charge density $\widetilde{\rho}_{\mathrm{m}}$ and magnetic current density $\widetilde{\boldsymbol{j}}_{\mathrm{m}}$, but no charged particles. It is interesting to speculate, therefore, on the possibility of a theory in which there are both charged particles and magnetic monopoles. In that case, Maxwell's equations would be

$$\begin{gathered}\nabla \cdot \boldsymbol{E} = \rho_{\mathrm{e}} \qquad \nabla \cdot \boldsymbol{B} = \rho_{\mathrm{m}} \\ \nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = -\boldsymbol{j}_{\mathrm{m}} \qquad \nabla \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{j}_{\mathrm{e}}.\end{gathered} \tag{13.80}$$

For this extended theory, the duality transformation (13.78) supplemented by the new change of variables

$$\widetilde{\rho}_{\mathrm{e}} = \rho_{\mathrm{m}} \qquad \widetilde{\boldsymbol{j}}_{\mathrm{e}} = \boldsymbol{j}_{\mathrm{m}} \tag{13.81}$$

leaves the form of the equations exactly the same: the theory is *self-dual*.

At the classical level, the extended Maxwell equations (13.80) are perfectly consistent with each other and (as readers may easily check) with the equations of continuity that express the conservation of both electric and magnetic charge, so there seems to be no fundamental reason why magnetic monopoles should not exist. Whether these equations make sense as part of a quantum-mechanical theory is quite another matter. In fact, we can see an immediate difficulty. Since $\nabla \cdot \boldsymbol{B}$ is not equal to zero, we can no longer express the magnetic field as $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. The whole construction of QED in chapters 8 and 9 was based on the possibility of expressing electromagnetic fields in terms of the 4-vector potential A_μ and would need serious rethinking if it were to accommodate the extended Maxwell equations.

A limited step in this direction was taken by Dirac, who considered the implications for an electrically-charged quantum-mechanical particle of the existence of a *classical* magnetic field produced by a magnetic monopole. The only equations we know of for the wavefunction or field operator for a charged particle, such as (8.13) for a spin-$\frac{1}{2}$ particle or the relativistic version of (13.63) for a spin-0 particle, involve the vector potential $\boldsymbol{A}$ rather than $\boldsymbol{B}$. Dirac's theory retains these equations, but allows for some modification in the relation between $\boldsymbol{B}$ and $\boldsymbol{A}$. For the moment, let us write $\boldsymbol{B} = \nabla \times \boldsymbol{A} + \ldots$, leaving open the question of what has to be added. A monopole of magnetic charge g ought to produce a magnetic field

$$\boldsymbol{B}(\boldsymbol{x}) = -\frac{g}{4\pi}\nabla\left(\frac{1}{r}\right) = \frac{g}{4\pi r^3}(x, y, z) \tag{13.82}$$

where $r = (x^2 + y^2 + z^2)^{1/2}$. To begin constructing this, consider the vector potential

$$\boldsymbol{A}(\boldsymbol{x}) = -\frac{g}{4\pi r(r-z)}(-y, x, 0). \tag{13.83}$$

By differentiating this expression, we find that $\nabla \times \boldsymbol{A}$ is equal to the function (13.82), but there is a catch. The vector potential (13.83) is singular at $r = z$, which means everywhere along the positive z axis $x = y = 0$, $z > 0$. Thus, there may be an additional contribution to $\nabla \times \boldsymbol{A}$ proportional to $\theta(z)\delta(x)\delta(y)\hat{\boldsymbol{z}}$, where $\hat{\boldsymbol{z}}$ is a unit vector in the positive z direction.

We can investigate this possibility by applying Stokes' theorem to the small cap S, shown in figure 13.4, on a sphere of radius r. The line element on the curve C is $\mathrm{d}\boldsymbol{\ell} = r\sin\theta(-\sin\varphi, \cos\varphi, 0)\mathrm{d}\varphi$ so, converting (13.83) to polar coordinates, we get

$$\int_S \nabla \times \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{S} = \oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{\ell} = -\frac{g}{4\pi}\left(\frac{\sin^2\theta}{1-\cos\theta}\right)\int_0^{2\pi}\mathrm{d}\varphi = -\frac{g}{2}\left(\frac{\sin^2\theta}{1-\cos\theta}\right). \tag{13.84}$$

When we shrink the cap to a point by taking $\theta \to 0$, the integral is just equal to $-g$. This non-zero value can come only from integrating the expression

$$\nabla \times \boldsymbol{A} = \frac{g}{4\pi r^3}(x, y, z) - g\theta(z)\delta(x)\delta(y)\hat{\boldsymbol{z}} \tag{13.85}$$

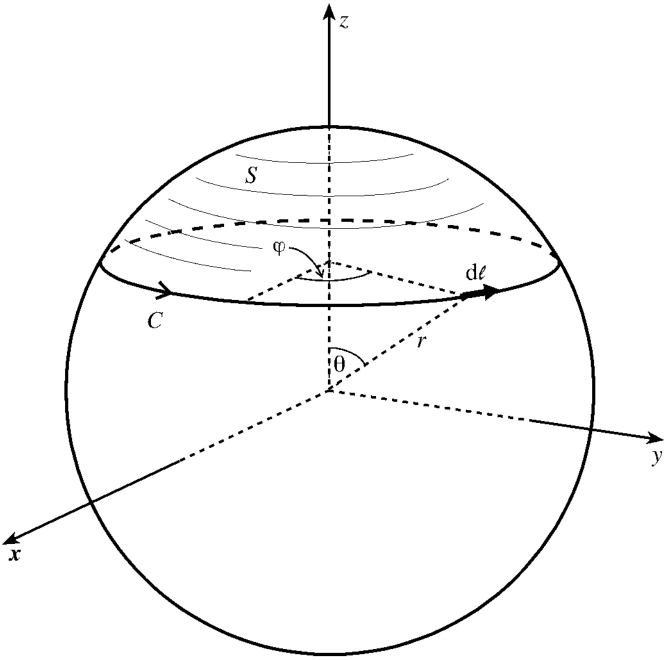


Figure 13.4. Spherical surface surrounding a Dirac monopole at the origin. As calculated from (13.84), the magnetic flux through the cap S remains non-zero when S is shrunk to an infinitesimal disc around the z axis.

the integral of the first term over an infinitesimal surface being zero. One can think of $\nabla \times \boldsymbol{A}$ as representing the magnetic field produced by an infinitely thin solenoid situated on the positive z axis between $z = 0$ and $z = \infty$; the first term is the field emerging from its open end at the origin, while the second is the field in its core. If, on the other hand, we take $\theta \to \pi$, so that S becomes the whole spherical surface, then the integral (13.84) vanishes, so the flux in the core of the 'solenoid' exactly cancels the flux passing through the rest of the spherical surface. Clearly, the field we actually want is

$$\boldsymbol{B}(\boldsymbol{x}) = \nabla \times \boldsymbol{A}(\boldsymbol{x}) + g\theta(z)\delta(x)\delta(y)\hat{\boldsymbol{z}}. \qquad (13.86)$$

The 'solenoid' is usually called the 'Dirac string'. It has little to do with the strings of the last section, and is in fact completely unobservable, as we can see in the following way. The same magnetic field can be obtained from the vector potential

$$\boldsymbol{A}'(\boldsymbol{x}) = \frac{g}{4\pi r(r+z)}(-y, x, 0) \qquad (13.87)$$

using

$$\nabla \times \boldsymbol{A}' = \frac{g}{4\pi r^3}(x, y, z) + g\theta(-z)\delta(x)\delta(y)\hat{z} \tag{13.88}$$

$$\boldsymbol{B}(\boldsymbol{x}) = \nabla \times \boldsymbol{A}'(\boldsymbol{x}) - g\theta(-z)\delta(x)\delta(y)\hat{z} \tag{13.89}$$

the Dirac string now occupying the negative z axis. These two vector potentials are related by a gauge transformation

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{x}) &= \boldsymbol{A}'(\boldsymbol{x}) - \frac{g}{2\pi(x^2+y^2)}(-y, x, 0) = \boldsymbol{A}'(\boldsymbol{x}) - \nabla\Theta(\boldsymbol{x}) \\ \Theta(\boldsymbol{x}) &= \frac{g}{2\pi}\tan^{-1}(y/x) = \frac{g}{2\pi}\varphi \end{aligned} \tag{13.90}$$

so we can regard $\boldsymbol{B}$ as being gauge invariant, provided that we adopt the rule of adjusting the Dirac string term in going from (13.86) to (13.89) as well as using the new vector potential. In fact, by making a suitable gauge transformation of this kind, we can make the string occupy any continuous curve (not necessarily a straight line) from the origin to infinity, so the string is an unphysical gauge degree of freedom.

Perhaps the most satisfactory way of dealing with these singular vector potentials is that devised by T T Wu and C N Yang. We divide space into two overlapping regions, say R and R', such that R does not include the positive z axis and R' does not include the negative z axis. (Strictly speaking, this means that neither R nor R' can include the origin, so we have to regard the point where the monopole is situated as being excluded from our space.) Then the quantum state of, say, a spin-$\frac{1}{2}$ particle in the presence of the monopole is described by a pair of wavefunctions, $\psi(x)$ which exists only in R and $\psi'(x)$ which exists only in R'. If the particle has an electric charge e, then they obey the equations

$$[\mathrm{i}\gamma^\mu\partial_\mu - e\gamma^\mu A_\mu(x) - m]\psi(x) = 0 \qquad \text{valid in } R \tag{13.91}$$

$$[\mathrm{i}\gamma^\mu\partial_\mu - e\gamma^\mu A'_\mu(x) - m]\psi'(x) = 0 \qquad \text{valid in } R' \tag{13.92}$$

where, to describe a static monopole, we can take $A_0(x) = 0$. Thus, neither wavefunction ever meets a string. At this point, we come to the central result of Dirac's theory. If $\psi(x)$ and $\psi'(x)$ together are to describe a unique state, then in the region $R \cap R'$ where R and R' overlap they must be related by the gauge transformation (13.90), which is to say

$$\psi'(x) = \exp[\mathrm{i}e\Theta(x)]\psi(x). \tag{13.93}$$

However we choose R and R', it will always be possible to find a closed curve in $R \cap R'$ which encircles the z axis. If we take the point x once round such a curve, then φ changes by $\pm 2\pi$ and $\Theta(x)$ changes by $\pm g$. Since both $\psi(x)$ and $\psi'(x)$ must be single-valued, this implies that

$$eg = 2\pi n \tag{13.94}$$

where n is a positive or negative integer. This result, the *Dirac quantization condition*, is a rather startling one. The wavefunction for any particle exists, in principle, throughout the universe, even though it may be exceedingly small outside of some localized region. Therefore, if one monopole of strength g exists anywhere in the universe, then the electric charge of every particle in the universe must be some multiple of $2\pi/g$.

These considerations do not constitute a comprehensive theory of magnetic monopoles. In particular, we simply assumed the existence of the magnetic field (13.82); we have no dynamical theory of any objects analogous to the charged particles of standard QED that might produce this magnetic field. In some non-Abelian gauge theories, the situation is quite different. Objects having the properties of magnetic monopoles can arise as soliton solutions to the equations of motion, without the need for *ad hoc* additions to the existing theory. The topological requirement for the existence of monopoles should be fairly obvious from our earlier discussions. If a Higgs field $\phi(\boldsymbol{x})$ is to have a finite energy, then its value everywhere on the surface of a large sphere (topologically, a two-sphere S^2), say with $|\boldsymbol{x}| = R$, must approach a point on the vacuum manifold as $R \to \infty$. The set of values that ϕ takes on over this spherical surface lie on a two-sphere drawn in the vacuum manifold. A topologically stable soliton can exist if it is possible to draw a two-sphere on the vacuum manifold which cannot be shrunk to a point. This is just the three-dimensional version of the criterion for the existence of vortices in two dimensions. Just as the simplest theory with vortices is one whose vacuum manifold is itself a circle, so the simplest possibility for a theory with monopoles is one whose vacuum manifold is itself a two-sphere. In fact, the three-dimensional analogue of the model (13.55) is one in which ϕ is a vector with three components, living in a three-dimensional internal space such as the isospin space of an SU(2) gauge theory.

The simplest non-Abelian magnetic monopole, which has come to serve as a standard pedagogical example, was discovered by G 't Hooft and A M Polyakov. It occurs in the SU(2) gauge theory whose Lagrangian density is

$$\mathcal{L} = -\tfrac{1}{4}G^a_{\mu\nu}G^{a\mu\nu} + \tfrac{1}{2}(\mathrm{D}_\mu\phi^a)(\mathrm{D}^\mu\phi^a) - \tfrac{1}{4}\lambda(\phi^a\phi^a - v^2)^2 \tag{13.95}$$

where

$$G^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu - e\epsilon^{abc}W^b_\mu W^c_\nu \tag{13.96}$$

is the field strength tensor for an isospin triplet of gauge fields W^a_μ (which I denote by $G^a_{\mu\nu}$ to distinguish it from the electromagnetic tensor $F_{\mu\nu}$) and e is the coupling strength, which can be identified with a fundamental electric charge. In contrast to the GWS model of §12.2, the Higgs field is an isospin triplet, whose gauge-covariant derivative is

$$\mathrm{D}_\mu\phi^a = \partial_\mu\phi^a - e\epsilon^{abc}W^b_\mu\phi^c. \tag{13.97}$$

This field theory is sometimes called the Georgi–Glashow model, because it was studied by H Georgi and S L Glashow as a possible (though ultimately unsatisfactory) model of electroweak interactions. If the Higgs field has a *constant* expectation value $\langle 0|\phi^a|0\rangle = (0, 0, v)$, then the terms in (13.95) which give masses to the gauge bosons are

$$\tfrac{1}{2}e^2v^2\epsilon^{ab3}\epsilon^{ac3}W^b_\mu W^{c\mu} = \tfrac{1}{2}e^2v^2\left(W^1_\mu W^{1\mu} + W^2_\mu W^{2\mu}\right). \tag{13.98}$$

The particles created by W^3_μ, the gauge field corresponding to the direction of $\langle 0|\phi|0\rangle$ in isospin space, remain massless. Within this model, they can be identified as photons, so the electromagnetic field is $A_\mu = W^3_\mu$.

The monopole is a configuration in which all the fields are static (that is, they are independent of t). With a suitable choice of gauge, we can also take $W^a_0 = 0$. In that case the energy density is equal to $-\mathcal{L}$ and we find

$$\mathcal{E} = \tfrac{1}{4}G^a_{ij}G^{aij} + \tfrac{1}{2}(\mathrm{D}_i\phi^a)(\mathrm{D}_i\phi^a) + \tfrac{1}{4}\lambda(\phi^a\phi^a - v^2)^2. \tag{13.99}$$

For a monopole centred at the origin, we take

$$\phi^a(\boldsymbol{x}) = v\frac{x^a}{r}\rho(r) \qquad W^a_i = -\epsilon_{aij}x^j f(r). \tag{13.100}$$

The symmetry of this trial solution is such that the Euler–Lagrange equations reduce to just two equations for the functions $\rho(r)$ and $f(r)$. As in the two-dimensional models, appropriate solutions exist, but they cannot be found exactly. The form in which the fields are expressed relies on the possibility of setting up a correspondence between directions in isospin space and directions in real space. As in figure 13.2, we can visualize the vector $\boldsymbol{\phi}(\boldsymbol{x})$ in isospin space as equivalent to a vector in real space attached to the point $\boldsymbol{x}$. For the configuration (13.100), the three quantities x^a/r are the components of a unit vector pointing radially outwards, so this is the three-dimensional analogue of the $n = 1$ configuration in figure 13.2. Here, the winding number means the number of times that the vacuum manifold is covered by the values of ϕ on the spherical surface at infinity. It is given by an expression analogous to (13.20) or (13.61), namely

$$n_S = (8\pi)^{-1}\int_V \mathrm{d}^3x\,\nabla\cdot\boldsymbol{u} = (8\pi)^{-1}\int_S \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{S} \tag{13.101}$$

where V is the volume enclosed by a closed surface S and $\boldsymbol{u}$ is the vector field whose components are

$$u^i = \epsilon^{ijk}\epsilon^{abc}\hat{\phi}^a(\partial_j\hat{\phi}^b)(\partial_k\hat{\phi}^c) \tag{13.102}$$

$\hat{\boldsymbol{\phi}}(\boldsymbol{x})$ being a unit vector in the direction of $\boldsymbol{\phi}(\boldsymbol{x})$. In our case, this unit vector is $\hat{\phi}^a = x^a/r$. Readers should not find it hard to verify that $u^i = 2x^i/r$ and hence that $n_S = 1$ when S is a sphere centred on the origin. (The expression for u^i is greatly simplified by the fact that $\epsilon^{abc}x^ax^b = \epsilon^{abc}x^ax^c = 0$, on account of the antisymmetry of ϵ^{abc}.)

The essential properties of the monopole can be deduced without knowing the detailed form of the functions $\rho(r)$ and $f(r)$, from the requirement that its energy be finite. First of all, we must have $\rho(r) \approx 1$ when r is large, so that the integral over all space of the potential energy term in (13.99) is finite. Taking $\rho(r) = 1$, we find that when r is large, the covariant derivative (13.97) is

$$\begin{aligned} \mathrm{D}_i\phi^a &\approx v\left(\frac{\delta^{ia}}{r} - \frac{x^i x^a}{r^3} + e\epsilon^{abc}\epsilon_{bij}\frac{x^j x^c}{r}f(r)\right) \\ &\approx v\left(\frac{\delta^{ia}}{r} - \frac{x^i x^a}{r^3}\right)\left[1 - er^2 f(r)\right]. \end{aligned} \tag{13.103}$$

(Readers who wish to verify this will find the result of exercise 13.7(a) helpful.) The integral of $\frac{1}{2}(\mathrm{D}_i\phi^a)(\mathrm{D}_i\phi^a)$ in (13.99) must also be finite, so when r is large, we must have $\mathrm{D}_i\phi^a \to 0$ and therefore $f(r) \approx 1/er^2$. Thus, when r is large, the gauge fields are given approximately by

$$W_i^a = -\epsilon_{aij}\frac{x^j}{er^2}. \tag{13.104}$$

A tricky point is that we can no longer identify the electromagnetic field as $A_\mu = W_\mu^3$, because $\langle\phi^a\rangle$ points in different directions in different regions of space. A general means of identifying the electromagnetic field strength tensor $F_{\mu\nu}$ has been discussed by 't Hooft (1974), but for our purposes, the right answer in the region where r is large is given by taking the component of $G_{\mu\nu}^a$ in the direction of ϕ^a. That is

$$F_{ij}(\boldsymbol{x}) \approx \hat{\phi}^a(\boldsymbol{x})G_{ij}^a(\boldsymbol{x}) \approx \frac{\epsilon_{aij}x^a}{er^3}. \tag{13.105}$$

(Readers who wish to verify the final expression will find the result of exercise 13.7(b) helpful.) On comparing this with (3.51), and noting that $F^{ij} = F_{ij}$ for the spatial components, we find that the magnetic field at large distances from the monopole is

$$B^i \approx -x^i/er^3 \tag{13.106}$$

and this agrees with (13.82) provided that we identify the magnetic charge as

$$g = -4\pi/e. \tag{13.107}$$

This result is reminiscent of the Dirac quantization condition (13.94) if we take $n = -2$. It does not mean quite the same thing, though, because e and n in (13.94) refer to the charge of a particle moving in the monopole field and to the phase of its wavefunction—considerations which have played no role in our treatment of the non-Abelian monopole. It is also worth noting that the gauge and Higgs fields of the non-Abelian monopole are non-singular, and there is nothing analogous to the Dirac string. This is possible because the magnetic field is a combined effect of three gauge fields and three Higgs fields, rather than of a single vector potential.

In our study of the sine–Gordon theory, we saw that (i) the solitons of that theory could be reinterpreted as the particles of another theory, the massive Thirring theory; (ii) these particles are fermions, despite the fact that the sine–Gordon theory contains only a bosonic field; (iii) weak coupling in one theory corresponds to strong coupling in the other. We might well wonder to what extent these features also occur in more realistic theories, such as the Georgi–Glashow model. The third feature is evidently analogous to the relation (13.107): a small electric charge implies a large magnetic charge for the monopole and *vice versa*. As for the spin of a monopole, the 't Hooft–Polyakov monopole is a spherically symmetric object and has no spin. However, monopoles of other kinds are possible which do have spin, and it is again possible to make spin-$\frac{1}{2}$ objects from purely bosonic fields. This is because of the correlation between directions in space and directions in the internal isospin space that characterize the monopole solutions. A rotation of the monopole in space must be accompanied by a corresponding rotation in isospin space. If our model contains, say, a doublet of scalar fields with isospin $\frac{1}{2}$, then the transformation of the monopole under this combined rotation may correspond to that of a spin-$\frac{1}{2}$ object (see, for example, Jackiw (1977)). It seems that these monopoles will also behave as fermions or bosons, according to the requirements of the spin-statistics theorem, but this question cannot be settled in as definitive a manner as was possible for the sine–Gordon theory.

Finally, there is the intriguing possibility that a gauge theory with monopoles might be related by the idea of duality, with which this section began, to another gauge theory in which the roles of electrically charged particles and magnetically charged monopoles were reversed. This would be analogous to the duality between the sine–Gordon and massive Thirring theories. A duality of this kind would have far-reaching consequences, because a strong coupling in one theory, which makes calculations very difficult, corresponds to a weak coupling in the dual theory, where perturbation theory can be used to good effect. It does not seem that the mere existence of monopoles is sufficient to make this idea work and to the best of my knowledge it is not possible to prove that any two gauge theories really are dual to each other in this sense. Nevertheless, there is strong circumstantial evidence for duality in certain supersymmetric gauge theories, which can be exploited to obtain exact, nonperturbative information about the quantum-mechanical properties of these theories. This is a rather technical subject. I know of no elementary account of the ideas that are involved, but interested readers may like to consult Giveon and Kutasov (1999), Intriligator and Seiberg (1996) and the references supplied by these authors.

Exercises

13.1. Verify the orthonormality properties (13.14)–(13.16), using the standard integrals

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{sech}^2 x = 2 \qquad \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{sech}^4 x = \frac{4}{3}$$

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \cos(qx)\mathrm{sech}^2 x = \frac{\pi q}{\sinh(\frac{1}{2}\pi q)}$$

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \cos(qx)\mathrm{sech}^4 x = \frac{2\pi q(1+\frac{1}{4}q^2)}{3\sinh(\frac{1}{2}\pi q)}.$$

Other integrals you will need can be obtained from integrations by parts. In the case of (13.15), you will also need the Fourier representation of the Dirac δ function given in appendix A. You will find it advantageous to express $\tanh^2 x$ as $1 - \mathrm{sech}^2 x$ wherever possible, and may like to be warned that the algebra is quite lengthy!

13.2. For the two-dimensional field theory of §13.1, consider the restricted theory in which there is a static kink and no free mesons, by writing $\phi(x,t) = \phi_{\mathrm{K}}(x) + c_1(t) f_1(x)$. Show that the Hamiltonian for this simplified theory consists of a constant (the energy of the kink) plus the Hamiltonian for an harmonic oscillator of frequency $\omega_1 = \sqrt{3}m/2$. From the Euler–Lagrange equation (or Hamilton's equations), verify that $c_1(t)$ has the form shown in (13.35) and that a_1 and $a_1^{\dagger}$ have the commutation relations appropriate for operators that create and annihilate bound mesons.

13.3. (a) Observe that in one dimension, equation (13.1) is equivalent to the equation of motion for a Newtonian particle whose position is ϕ and whose potential energy is $-V(\phi)$, if x is taken to represent time. Sketch this potential energy, which should show two 'hills' at $\phi = \pm v$. Convince yourself that the kink solution $\phi = \phi_{\mathrm{K}}$ and the 'anti-kink' solution $\phi = -\phi_{\mathrm{K}}$ correspond to this particle's being infinitesimally displaced from the top of one hill at 'time' $x = -\infty$ and eventually coming to rest at the top of the other hill at 'time' $x = +\infty$.

(b) Convince yourself that there are further solutions (which are hard to write down in closed form), consisting of an alternating sequence of kinks and anti-kinks, in which the analogue Newtonian particle spends most of its 'time' moving very slowly near the hilltops and brief intervals of 'time' traversing the valley. (Strictly speaking, these kinks and anti-kinks must be infinitely far apart: if the solutions are to have a finite energy, the analogue particle must come infinitesimally close to the top of each hill, where its 'speed' is infinitesimal. In effect, only single-kink and single-anti-kink solutions to the time-independent equation (13.1) are allowed. However, the time-dependent equation (13.25) does have allowed solutions consisting of sequences of *moving* kinks and anti-kinks, separated by finite distances. Kink–anti-kink pairs will eventually collide, however, and may well disappear, their energy being converted into mesons. The restricted definition of a soliton mentioned in the text requires that true solitons should survive such collisions intact.)

(c) Sketch a function $\phi(x)$ corresponding to a kink and an anti-kink, and a second function in which the positions of the kink and anti-kink are interchanged. Convince yourself that kinks and anti-kinks are fermions.

13.4. For the purposes of this exercise, let us denote the operators of the quantum sine-Gordon theory by $\hat{\phi}_1(x', t)$ and $\hat{\psi}_1(x, t)$ to distinguish them from ordinary functions. Suppose that $|\phi\rangle$ is an eigenstate of $\hat{\phi}(x', t)$ with eigenvalue $\phi(x')$. Following the method of (5.62), use the commutator (13.51) to show that $\hat{\psi}_1(x, t)|\phi\rangle$ is an eigenstate of $\hat{\phi}(x', t)$ with eigenvalue $\phi(x') + \Delta\phi(x')$, where $\Delta\phi(x')$ is given by (13.53).

13.5. Assume that when r is large the solution to (13.58) is given approximately by $\rho \approx v + c_n r^{-p}$. By substituting this trial solution, show that $p = 2$ and $c_n = -n^2 v/m^2$.

13.6. In two spatial dimensions, let $r = \sqrt{x^2 + y^2}$. Verify that $\nabla^2 \ln(r/a) = 0$, except at $r = 0$, where the answer is not well defined. In two dimensions, Gauss' theorem is

$$\int_S \nabla \cdot \boldsymbol{v}\, \mathrm{d}^2 x = \oint_C \boldsymbol{v} \cdot \mathrm{d}\boldsymbol{\ell}$$

where S is the area bounded by a closed curve C. Taking $\boldsymbol{v} = \nabla \ln(r/a)$ and C to be any circle centred on the origin, show that $\int_S \nabla^2 \ln(r/a)\, \mathrm{d}^2 x = 2\pi$, and hence that $\nabla^2 \ln(r/a) = 2\pi\delta(\boldsymbol{x})$. Note that a is an arbitrary length, needed to make the argument of the logarithm dimensionless. Since $\ln(r/a') = \ln(r/a) + \ln(a/a')$, a change in this arbitrary length is equivalent to adding a constant to the Coulomb potential. As in three dimensions, this constant has no physical meaning. However, the usual convention of taking the potential to vanish as $r \to \infty$ obviously doesn't work in two dimensions.

13.7. (a) The expression $\epsilon^{abc}\epsilon_{bij}x^j x^c$ defines a 3-dimensional tensor with two indices, i and a, and is quadratic in the x^k. It must be of the form $Ar^2\delta^{ia} + Bx^i x^a$, where A and B are constants. By considering the case $i = a = 1$, show that $A = -1$ and $B = 1$, and check that the result is also true for some other values of i and a.

(b) The quantity $\epsilon^{abc}\epsilon_{bik}\epsilon_{cjl}x^k x^l$ is also quadratic in the x^k, but it has three indices a, i and j. Convince yourself that it is antisymmetric in the indices i and j, and must be of the form $A'r^2\epsilon_{aij} + B'x^a x^k \epsilon_{kij}$ where A' and B' are constants. By considering the case $a = 1$, $i = 2$, $j = 3$, show that $A' = 0$ and $B' = 1$, and check that the result is also true for some other values of a, i and j.

Chapter 14

The Early Universe

In this chapter, I shall discuss an area of investigation that illustrates many of the theoretical ideas developed in the rest of the book, namely cosmology and the early history of the universe. Only in the last 60 years or so has it been possible to treat cosmology as a matter for serious scientific enquiry rather than philosophical speculation. Since we cannot (presumably) create a new universe in the laboratory, any theory concerning the history of our own universe must remain to some extent speculative. If, however, it is accepted that our knowledge of physics as established in the laboratory and by astronomical observations continues to be valid in the distant past, then a remarkable amount can be said with a fair degree of confidence. For example, the present age of the universe is known, roughly to within a factor of 2: it cannot be much less than 10 billion years (1 billion $= 10^9$) nor much greater than 20 billion years. Our established knowledge of physics can, of course, be applied with confidence only when conditions in the universe were such that a confident extrapolation can be made from conditions which can be created in the laboratory. This has been true ever since the universe was about one millisecond old. In the first millisecond, however, events moved extremely rapidly.

As we shall see, the temperature of the matter in the universe increases, without any known limit, as we progress backwards in time, and our reasoning about what the sequence of events may have been becomes increasingly speculative as we encounter energies at which our confidence in the standard model of particle physics begins to falter. Conversely, it is potentially fruitful to speculate about early events on the basis of theoretical models, such as grand unified theories, which cannot be rigorously tested in the laboratory. The reason is that the very early cosmological events implied by these models may have consequences for the present constitution of the universe that can be checked by astronomers. This opens the enticing possibility of using the early universe as a high-energy physics laboratory in which energies are available that could not conceivably be produced by man. Some fragments of information have already been obtained in this way. Clearly, however, the reliability of such information is no greater than the reliability with which the detailed consequences of theoretical models can be worked out. At present, there is, in my view at least, little cause for complacency in this respect.

I shall begin by outlining the standard *big bang* model of the history of the universe.

14.1 The Robertson–Walker Metric

Modern cosmology is based upon the description of spacetime geometry given by general relativity. As I have mentioned from time to time, there is a widespread belief that general relativity is inadequate as a fundamental theory of spacetime geometry, to the extent that it is non-quantum-mechanical. If this is so, then there is a limit to the validity of the standard cosmological model, which I shall discuss in due course. For the moment, let us assume that general relativity is good enough. We need to write down the metric tensor of the universe. Obviously, it is impossible to do this in any detail, but by a happy circumstance, astronomical evidence shows that the overall structure of that part of the universe which can be observed is very simple. If the distribution of matter is averaged over distances that are large enough to encompass many clusters of galaxies, it is found to be *isotropic*, which means that it looks the same in all directions, and *homogeneous*, which means that it would look the same from any vantage point. The best evidence for isotropy actually comes from measurements of the cosmic microwave radiation, which we shall have cause to discuss later on. Our first basic assumption, then, is that the universe is isotropic and homogeneous. This assumption is sometimes dignified as the *cosmological principle*. It can be seen as embodying the philosophical prejudice that our own location in the universe has nothing whatever to distinguish it from any other location. Its only scientific value, however, is that it is in reasonable accord with observations and that it makes further progress possible.

From the assumptions of homogeneity and isotropy, it can be shown to follow that there is a coordinate system in which the line element (2.7) has the form (with $c = 1$)

$$\mathrm{d}\tau^2 = \mathrm{d}t^2 - a^2(t)\left(\frac{1}{1-kr^2}\,\mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right). \tag{14.1}$$

This is called the *Robertson–Walker* line element. The second term, in which k is a constant, measures distances in a spatial section of the spacetime, which exists at an instant t of *cosmic time*. The physical distance between two points in this space separated by fixed coordinate intervals $\mathrm{d}r$, $\mathrm{d}\theta$ and $\mathrm{d}\phi$ varies with time in proportion to the function $a(t)$, called the *scale factor*, which depends only on time. As in the Schwarzschild line element (4.28), the coordinate r does not necessarily provide a linear measure of distance. However, t does measure a genuine time. The proper time τ measured by any observer whose

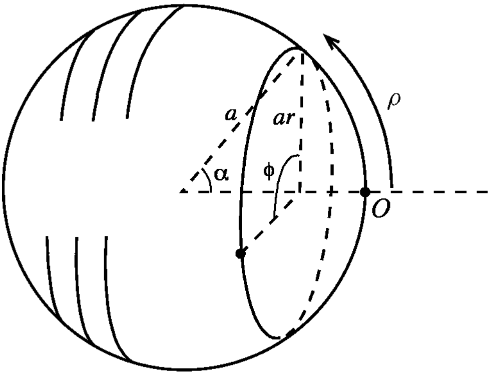


Figure 14.1. The surface of a sphere of radius a represents two of the spatial dimensions of a closed Robertson–Walker universe. The volumes inside and outside the surface are not part of the space. Relative to the origin at O, the coordinates r and ϕ can be visualized as shown. The physical distance from O to a point on the circle of radius ar in the figure is ρ. In the full three-dimensional spatial section, ρ would be the physical radius of a sphere centred at O.

spatial coordinates r, θ and ϕ are fixed is clearly the same as t. Moreover, such an observer is moving through the spacetime along a geodesic and is therefore in free fall, which would not be the case in the Schwarzschild spacetime. (It would be a good exercise for readers to verify this point by deriving the geodesic equations, using the method suggested by (4.27).) The sequence of spatial sections corresponding to successive instants of cosmic time can be thought of as a three-dimensional space that expands or contracts uniformly with time according to the variation of $a(t)$. The surfaces of constant r, θ or ϕ expand or contract in the same way, like a grid of lines painted on the surface of an inflating balloon, and these coordinates are said to be *comoving*.

The constant k in (14.1) may be positive, negative or zero. If it is non-zero, then we can make the change of variables $r \to r/|k|^{1/2}$ and $a(t) \to a(t)|k|^{1/2}$, so that the magnitude of k disappears. We can therefore always choose the coordinates so that k has one of the three values 1, 0 or -1. If $k = 0$, then the spatial part of (14.1) is just the line element of a three-dimensional Euclidean space and the universe is *flat*. To understand the spatial geometry when $k = 1$, consider the two-dimensional surface $\theta = \pi/2$. The three-dimensional space can be thought of as the volume of revolution of this surface. The surface is in fact the surface of a Euclidean sphere of radius $a(t)$, as sketched in figure 14.1. In terms of the angles α and ϕ, the element of length $\mathrm{d}s$ on this surface is clearly given by $\mathrm{d}s^2 = a^2(\mathrm{d}\alpha^2 + \sin^2\alpha\,\mathrm{d}\phi^2)$, and this reproduces the spatial part of the Robertson–Walker line element when r is identified as $\sin\alpha$, as shown, for then we have $\mathrm{d}\alpha = \mathrm{d}\sin^{-1} r = \mathrm{d}r/(1 - r^2)^{1/2}$. It will be seen that a given value of r actually corresponds to two values of α, namely $\alpha = \sin^{-1} r$ and $\alpha = \pi - \sin^{-1} r$

so the coordinate r provides an unambiguous label only for points on one half of the sphere, say with $\alpha < \pi/2$ or $r < 1$. The singularity in (14.1) at $r = 1$ is only a coordinate singularity, which marks the edge of the region in which r is a valid coordinate. The spherical surface obviously *is* isotropic and homogeneous, and the origin $r = 0$ could be placed anywhere on it. At a given instant of time, the volumes inside and outside the spherical surface in figure 14.1 have nothing to do with the Robertson–Walker *space*, and serve only as an aid to visualizing the surface. On the other hand, the sequence of spatial sections that are obtained as $a(t)$ varies with time can be envisaged as a set of concentric spherical surfaces that fill all or part of this volume. Each spatial section can be described as having a (spatially) constant positive radius of curvature $a(t)$.

Consider now a sphere drawn in the Robertson–Walker space at fixed coordinate radius r. Its physical radius is

$$\rho(r) = a\int_0^r \frac{\mathrm{d}r}{(1-r^2)^{1/2}} = a\sin^{-1}(r) = a\alpha. \tag{14.2}$$

The circumference of a great circle drawn on this sphere, say the equator $\theta = \pi/2$, is

$$c(r) = a\int_0^{2\pi} r\,\mathrm{d}\theta = 2\pi a r = 2\pi a\sin(\rho/a) \tag{14.3}$$

which is always smaller than $2\pi\rho$, as is evident from figure 14.1. This circumference has a maximum value of $2\pi a$ at $\rho = \pi a/2$ and decreases to zero at $\rho = \pi a$. Thus, for $k = 1$, the spatial section of the Robertson–Walker universe is a three-dimensional spherical surface and is said to be *closed*.

For $k = -1$, the spatial section has a constant negative radius of curvature and is more difficult to imagine pictorially, though an analogy is often made with the surface of a saddle. The radius and circumference of a sphere are

$$\rho(r) = a\sinh^{-1}(r) \qquad \text{and} \qquad c(r) = 2\pi a\sinh(\rho/a). \tag{14.4}$$

The circumference is always greater than $2\pi\rho$ and both can be arbitrarily large. This universe has an infinite spatial extent and is said to be *open*.

We shall need to know the Ricci tensor, which appears in the field equations (4.17) of general relativity. The metric tensor, whose components appear in (14.1), is diagonal, with $g_{00} = 1$ and spatial components given by

$$g_{ij} = -a^2\begin{pmatrix} (1-kr^2)^{-1} & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2\theta \end{pmatrix}. \tag{14.5}$$

We find that the Ricci tensor is also diagonal, given by

$$R_{00} = -3\frac{\ddot{a}}{a} \qquad \text{and} \qquad R_{ij} = -\left(\frac{\ddot{a}}{a} + 2\frac{\dot{a}^2}{a^2} + 2\frac{k}{a^2}\right)g_{ij} \tag{14.6}$$

where the overdots stand for $\partial/\partial t$. The Ricci scalar is

$$R = g^{\mu\nu}R_{\mu\nu} = -6\left(\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right) \tag{14.7}$$

and the Einstein curvature tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$ is diagonal, with components given by

$$G_{00} = 3\left(\frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right) \qquad \text{and} \qquad G_{ij} = \left(2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right)g_{ij}. \tag{14.8}$$

If the metric of our universe is approximately of the Robertson–Walker form, and if the scale factor does change with time, then a simple consequence is *Hubble's law*. Assume that our galaxy and those we observe are comoving, so that their spatial coordinates are fixed. Then the physical distance between two galaxies separated by a coordinate distance d_0 is $d = a(t)d_0$. Their relative velocity is therefore

$$v = \frac{\mathrm{d}}{\mathrm{d}t}d(t) = \frac{\dot{a}}{a}d(t). \tag{14.9}$$

This velocity is proportional to the distance between the galaxies, with the proportionality factor

$$H(t) = \dot{a}(t)/a(t). \tag{14.10}$$

It is, of course, unlikely that galaxies will be exactly comoving. Nevertheless, it was discovered by E Hubble in 1929 that distant galaxies are, on average, receding from us with velocities proportional to their distances from us, so the universe is expanding. The velocity of recession can be measured as a redshift of spectral lines, and the distance in terms of the apparent luminosity of an object whose absolute luminosity is known. The redshift z is defined by $z = (\lambda_\mathrm{o}/\lambda_\mathrm{e}) - 1$, where λ_o is the observed wavelength and λ_e is the wavelength of light at the moment it was emitted, as it would be determined in the rest frame of the radiating object. When z is small, it can be interpreted as a non-relativistic Doppler shift. More generally, however, careful account must be taken of the change in $a(t)$ between the moments of emission and reception. The relation between luminosity distance d_L and redshift can be written as a power series (see exercise 14.1)

$$d_\mathrm{L} = H_0^{-1}\left[z + \tfrac{1}{2}(1 - q_0)z^2 + \ldots\right] \tag{14.11}$$

where the *Hubble constant* H_0 is the present value of $H(t)$ and q_0 is the present value of the *deceleration parameter*

$$q = -a\ddot{a}/\dot{a}^2. \tag{14.12}$$

The values of H_0 and q_0 are not known with very high precision. Hubble's constant is usually expressed as

$$H_0 = h \times 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}} = h(9.78 \times 10^9\,\mathrm{years})^{-1} \tag{14.13}$$

and the dimensionless number h deduced from observations is between 0.5 and 0.8. (Clearly, H has dimensions $(\text{time})^{-1}$, but the units in which it is traditionally measured are recessional velocity (km s^{-1}) per unit distance to a galaxy, measured in megaparsecs, with $1\,\text{Mpc} = 3.086 \times 10^{22}\,\text{m}$.) The value of q_0 is rather uncertain, because very distant galaxies must be observed to detect any curvature in the plot of d_{L} against z. It has generally been thought that the expansion rate of the universe must be slowing down, and thus that q_0 must be positive, but recent evidence suggests that this may not be so after all. I shall say a little more about this in the next section, where we shall be able to see more clearly what is at stake.

For many purposes, including the derivation of (14.11), it is necessary to understand the behaviour of light waves in the Robertson–Walker universe. Of course, the propagation of electromagnetic waves in general spacetimes can be investigated systematically by the methods we touched on in §7.7, but the essential fact pertaining to a Robertson–Walker universe can be discovered in a more elementary, and perhaps more enlightening way as follows. It will be sufficient to consider the case of a wave emitted by a comoving atom, say at $r = r_{\text{e}}$ and $\theta = \phi = 0$, and received by a comoving observer at $r = 0$. The light ray moves along a null geodesic whose equation, according to (14.1) is $\mathrm{d}t = -a(t)\mathrm{d}r/(1 - kr^2)^{1/2}$, the negative square root corresponding to a ray moving towards the origin. If a wave crest is emitted at time t_{e} and received at time t_{o}, then

$$\int_{t_{\text{e}}}^{t_{\text{o}}} \frac{\mathrm{d}t}{a(t)} = \int_0^{r_{\text{e}}} \frac{\mathrm{d}r}{(1 - kr^2)^{1/2}} = d_0 \tag{14.14}$$

where the coordinate distance d_0 travelled by the ray is independent of both t_{e} and t_{o}. If the following crest is emitted at $t_{\text{e}} + \Delta t_{\text{e}}$ and received at time $t_{\text{o}} + \Delta t_{\text{o}}$, then

$$\int_{t_{\text{e}}+\Delta t_{\text{e}}}^{t_{\text{o}}+\Delta t_{\text{o}}} \frac{\mathrm{d}t}{a(t)} = d_0 + \frac{\Delta t_{\text{o}}}{a(t_{\text{o}})} - \frac{\Delta t_{\text{e}}}{a(t_{\text{e}})} = d_0 \tag{14.15}$$

and so $\Delta t_{\text{o}}/\Delta t_{\text{e}} = a(t_{\text{o}})/a(t_{\text{e}})$. Thus, the observed frequency $\nu_{\text{o}} = 1/\Delta t_{\text{o}}$ and wavelength $\lambda_{\text{o}} = 1/\nu_{\text{o}}$ (in natural units, with $c = 1$) are related to those of the emitted wave by

$$\frac{\nu_{\text{o}}}{\nu_{\text{e}}} = \frac{a(t_{\text{e}})}{a(t_{\text{o}})} \qquad \text{or} \qquad \frac{\lambda_{\text{o}}}{\lambda_{\text{e}}} = \frac{a(t_{\text{o}})}{a(t_{\text{e}})}. \tag{14.16}$$

As seen by a comoving observer, therefore, the physical wavelength of a photon changes in proportion to the scale factor. In exercise 14.2, readers are invited to investigate this effect in terms of a covariant wave equation.

As we shall discuss later on, the universe is known to be filled with black-body radiation, the cosmic microwave background, whose current temperature is approximately 2.7 K. The result we have just obtained shows that the energy of a photon belonging to this background, and therefore the temperature of the photon gas as a whole, is proportional to $1/a(t)$.

14.2 The Friedmann–Lemaître Models

The Robertson–Walker metric on its own tells us nothing about the time dependence of the scale factor. To investigate this, we have to study the field equations (4.17), which involve the stress tensor for whatever matter is present. From the form of the metric tensor and the Einstein curvature tensor (14.8), it is clear that the stress tensor must be diagonal, with elements

$$T_{00} = \rho(t) \qquad \text{and} \qquad T_{ij} = -p(t)g_{ij} \tag{14.17}$$

where $\rho(t)$ and $p(t)$ are functions of time only. This is the only form of stress tensor that is consistent with the assumptions of isotropy and homogeneity. In a sufficiently small region, the metric must be approximately that of Minkowski spacetime and we can choose new spatial coordinates in which g_{ij} is diagonal, with each component equal to -1. Then, by comparing (14.17) with (3.43), we can identify ρ as the energy density and p as the pressure, provided that the matter behaves as a fluid in thermal equilibrium. The field equations now provide two independent equations relating $a(t)$, $\rho(t)$, $p(t)$ and the cosmological constant Λ, which are

$$3\left(\frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right) = \kappa\rho + \Lambda \tag{14.18}$$

$$2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -\kappa p + \Lambda. \tag{14.19}$$

By combining them, we find

$$\frac{\ddot{a}}{a} = -\frac{\kappa}{6}(\rho + 3p) + \frac{1}{3}\Lambda. \tag{14.20}$$

We saw in (4.24) that the quantity Λ/κ, which appears in (14.18) as an additional energy density and in (14.19) as an additional (negative) pressure, cannot be much greater than the average density of matter in the universe, which is extremely small compared with the densities of everyday materials. In general relativity, Λ is a fundamental constant, independent of the properties or distribution of any ordinary matter that the universe may happen to contain. To decide how large or small Λ is in a meaningful way, we must compare it with another fundamental quantity. The only constants at our disposal are G, $\hbar$ and c, and from these we can construct a quantity with the dimensions of a mass density for comparison with (4.24). It is $c^5/G^2\hbar \approx 5 \times 10^{93}\,\mathrm{g\,cm^{-3}}$. Thus, a dimensionless measure of the size of Λ is a number of the order of 10^{-122}. The staggering smallness of this number has led many theorists to suppose that the cosmological constant must be identically zero. Whether this number is really significant is hard to say. Since its derivation involves $\hbar$, any detailed understanding must require a reliable understanding of the relationship between spacetime geometry and quantum mechanics, which we do not have. A number of speculative arguments have been put forward to the effect that Λ should be zero, or at least very small, but none of them is conclusive. Suppose that Λ can be neglected in (14.20). For an ordinary fluid in thermal equilibrium, the density and pressure are both positive. Even if the matter is not in thermal equilibrium, the quantity $\rho + 3p$ is positive for ordinary matter under all ordinary circumstances, which is a special case of the *strong energy condition* discussed by Hawking and Ellis (1973). This being so, we see from (14.20) that $\ddot{a}$ is negative, which means that the rate of expansion of the universe should be decreasing, and so the deceleration parameter q_0 in (14.11) should be positive. As I mentioned in the last section, there is some evidence that q_0 is actually negative. It comes from the observation of supernovae (Perlmutter *et al*, 1998), which are bright enough (albeit short-lived) to permit estimates of the distances and redshifts of the very distant galaxies in which they occur. If the interpretation of these observations is correct, then the negative value of q_0 would indicate the existence of a small, positive cosmological constant. Although this possibility cannot be ignored in the ongoing process of constructing a detailed model of the universe, the effect of such a small cosmological constant on our overall cosmological picture would be relatively minor. For the purposes of this chapter, I propose to simplify matters by setting $\Lambda = 0$. The cosmological models based on the Robertson–Walker metric and Einstein's field equations with $\Lambda = 0$ are known as the *Friedmann–Lemaître models*.

It is convenient to write (14.18) and (14.19) with $\Lambda = 0$ as

$$\dot{a}^2 + k = \tfrac{1}{3}\kappa\rho a^2 \tag{14.21}$$

$$\ddot{a} = -\tfrac{1}{6}\kappa(\rho + 3p)a. \tag{14.22}$$

The first of these is sometimes referred to as the *Friedmann equation*. By differentiating it, we may easily show that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\rho a^3) = -p\frac{\mathrm{d}}{\mathrm{d}t}(a^3). \tag{14.23}$$

This equation is equivalent to $\nabla_\nu T^{\mu\nu} = 0$ and is usually said to express the conservation of energy. The physical volume V occupied by a given amount of matter is proportional to a^3, so if the internal energy of this matter is U, then (14.23) asserts that $\mathrm{d}U/\mathrm{d}t = -p\mathrm{d}V/\mathrm{d}t$. That is, the rate of change of U is equal to the rate at which work is done on the region in question by its surroundings.

To draw detailed conclusions from (14.21) and (14.22), we need information about ρ and p. Some general conclusions can be obtained without very detailed information, however. First, suppose that $k = 0$, so that the universe is flat. Then (14.21) gives a relation between the density and the Hubble parameter (14.10):

$$\rho(t) = \rho_{\mathrm{c}}(t) \equiv \frac{3}{\kappa}\frac{\dot{a}^2}{a^2} = \frac{3}{\kappa}H^2(t). \tag{14.24}$$

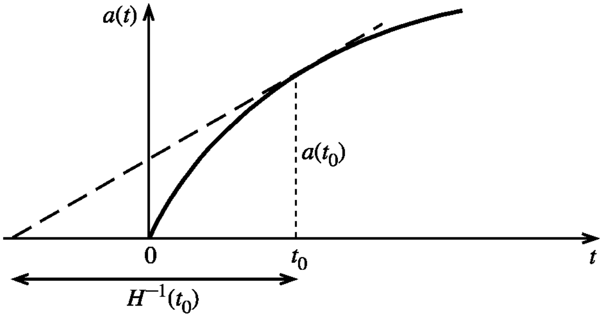


Figure 14.2. Scale factor of a Robertson–Walker universe in which the expansion rate always decreases. The age t_0 of the universe is less than $H^{-1}(t_0)$.

The quantity $\rho_c(t)$ is called the *critical density*. When k is not necessarily equal to zero, it is convenient to measure the density as a fraction of the critical density, defining

$$\Omega(t) = \rho(t)/\rho_c(t). \tag{14.25}$$

Then equation (14.21) becomes

$$\dot{a}^2(\Omega - 1) = k \tag{14.26}$$

and we see that in a closed universe, with $k = +1$, the density always exceeds the critical density ($\Omega > 1$), while in an open universe it is always less than the critical density ($\Omega < 1$).

According to our earlier discussion, (14.22) shows that $\ddot{a}$ is always negative, and therefore that $\dot{a}$ always decreases with time. (Readers may like to note the somewhat counter-intuitive result that a positive pressure acts to slow down, rather than to accelerate the expansion.) Since the universe is now observed to be expanding, the expansion rate increases as we look further back in time. It follows (see figure 14.2) that at some time in the past the scale factor a was equal to zero and that the time which has elapsed since then is less than $1/H_0$. When the scale factor is zero, the universe is infinitely compressed (although, if it is open or flat, its spatial extent is still infinite). This is a highly singular state, containing matter at infinite density. From a mathematical point of view, the metric becomes ill-defined, and the instant of time at which this occurs should be excluded from our spacetime manifold. Physically, we have no way of knowing what might happen in the extreme conditions prevailing near this singularity. From either point of view, the singularity marks the earliest time at which our universe can meaningfully be said to have existed. If we set $t = 0$ at the initial singularity, then the estimates of H_0 given above yield an upper bound to the present age of the universe t_0 of

$$t_0 < 2 \times 10^{10}\,\text{years}. \tag{14.27}$$

These conclusions are based on assumptions which could turn out to be false. The first was that the universe is homogeneous and isotropic, which is certainly not exactly true. We might wonder whether the occurrence of an initial singularity is a consequence of the high degree of symmetry, which might be avoided if allowance were made for anisotropies and inhomogeneities. It seems (as discussed, for example by Hawking and Ellis (1973)) that this is not so and that under quite general conditions an initial singularity must have occurred. On the other hand, the behaviour of the metric in the neighbourhood of the singularity may be much more complicated in an anisotropic and/or inhomogeneous universe than in the Friedmann-Lemaître models (see, for example, Misner *et al* (1973)). Another assumption was the strong energy condition $(\rho + 3p) > 0$. If this is not true, then there need not be an initial singularity because, going backwards in time, $\ddot{a}$ might become positive, allowing a to pass through a minimum and then increase. For ordinary matter, the strong energy condition holds. Later on, in connection with the *inflationary universe*, we shall encounter a situation in which the strong energy condition may cease to hold for a brief period of time, but this does not in itself avoid the initial singularity. Finally, the entire argument is based on a classical spacetime geometry. If, as is generally believed, this geometry is ultimately subject to quantum-mechanical laws, then we may expect these laws to become important when the universe is sufficiently small. Since we have no reliable quantum theory of gravity, it is not possible to be certain about when quantum effects will be important. A rough estimate can be obtained by requiring that the energy density should not exceed the characteristic value of $c^5/G^2\hbar$. At high densities, as we shall see shortly, the curvature term k/a^2 in (14.18) is negligible, even though a may be very small. Using this equation (and dimensional analysis to convert to laboratory units), we find that quantum gravity effects are likely to be important when

$$H^{-1} \lesssim \left(G\hbar c^{-5}\right)^{1/2} \approx 5 \times 10^{-44}\,\mathrm{s}. \tag{14.28}$$

Since H^{-1} is a rough measure of the age of the universe, this time, called the *Planck time*, is the time at which we expect that quantum gravity effects ceased to be important.

At the present time, it appears that the matter in the universe is fairly well described as a uniform, comoving distribution that exerts no pressure, known to cosmologists as *dust*. As a first approximation, it is instructive to suppose that this always has been and always will be true. Then the solution to (14.23) is

$$\rho(t) = M/a^3(t) \tag{14.29}$$

where M is a constant equal to the mass contained in a comoving region of physical volume $a^3(t)$. With $p = 0$, the deceleration parameter can be expressed as

$$q(t) = \frac{\kappa\rho(t)}{6H^2(t)} = \tfrac{1}{2}\Omega(t). \tag{14.30}$$

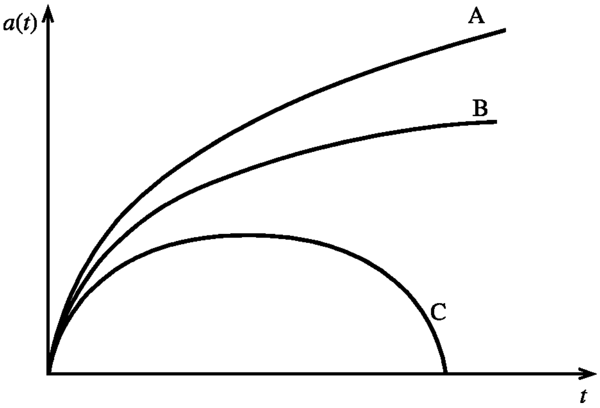


Figure 14.3. Variation of the scale factor with time in Friedmann–Robertson–Walker models: A, open universe, $k = -1$; B, flat universe, $k = 0$; C, closed universe, $k = 1$.

The variation of the scale factor with time can now be found by solving (14.21). For $k = 0$, the solution is

$$a(t) = \left(\tfrac{3}{4}\kappa M\right)^{1/3} t^{2/3}. \tag{14.31}$$

For $k = \pm 1$, it can be written in parametric form in terms of an angle θ:

$$a = \tfrac{1}{3}\kappa M \sin^2\theta \qquad t = \tfrac{1}{3}\kappa M\left(\theta - \tfrac{1}{2}\sin 2\theta\right) \qquad \text{for } k = 1 \tag{14.32}$$

$$a = \tfrac{1}{3}\kappa M \sinh^2\theta \qquad t = \tfrac{1}{3}\kappa M\left(\tfrac{1}{2}\sinh 2\theta - \theta\right) \qquad \text{for } k = -1. \tag{14.33}$$

These solutions are sketched in figure 14.3, and we see that both the open and flat universes continue to expand for ever, while the expansion of the closed universe eventually comes to a halt and this universe recollapses to a final singularity. The situation is quite analogous to that of a projectile launched from the Earth's surface, the flat universe corresponding to an initial velocity equal to the escape velocity (see exercise 14.4).

From the above solutions, it is possible to derive a relation between the age of the universe t, the Hubble parameter H and the density ratio Ω of the form

$$t = H^{-1} f(\Omega). \tag{14.34}$$

Since the open and closed universes correspond to $\Omega < 1$ and $\Omega > 1$ respectively, the function $f(\Omega)$ has different forms in these two ranges:

$$f(\Omega) = \begin{cases} \dfrac{1}{1-\Omega} - \dfrac{\Omega}{2}(1-\Omega)^{-3/2}\cosh^{-1}\left(\dfrac{2}{\Omega} - 1\right) & \text{for } \Omega \le 1 \\ \dfrac{\Omega}{2}(\Omega - 1)^{-3/2}\cos^{-1}\left(\dfrac{2}{\Omega} - 1\right) - \dfrac{1}{\Omega - 1} & \text{for } \Omega \ge 1. \end{cases} \tag{14.35}$$

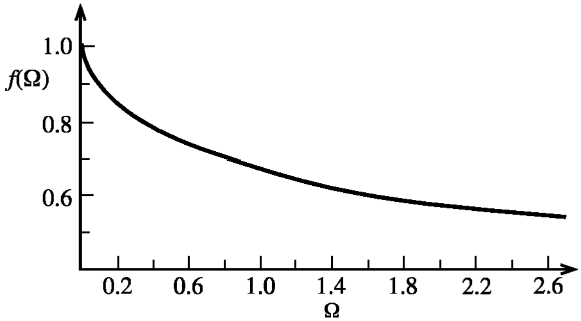


Figure 14.4. The function $f(\Omega)$ given in (14.35).

At $\Omega = 1$, both expressions reduce to $f(1) = \frac{2}{3}$, and $f(\Omega)$ is in fact a perfectly smooth function, plotted in figure 14.4.

14.3 Matter, Radiation and the Age of the Universe

From (14.34) and (14.35), we can determine the present age of the universe, provided that (i) we can assume that $p = 0$; (ii) we have an estimate of H_0; and (iii) we have an estimate of Ω_0, the present density as a fraction of the present critical density. The assumption that $p = 0$ is, for this purpose, perfectly safe. The period during which this has been true is called the *matter-dominated era* and, as we shall discover shortly, it began when the universe was about one millionth of its present age. The value of H_0 is, as we have seen, uncertain by something like a factor of 2, so the error in assuming that $p = 0$ is negligible by comparison. The value of Ω_0 is also rather uncertain. Direct observations reveal, of course, only *luminous* matter, namely that contained in stars whose radiation we can detect. There are, however, a number of reasons for believing that there is a considerable amount of additional matter, called *dark matter* or *missing matter*.

The masses of distant galaxies are estimated by means of the *virial theorem*, which asserts, roughly, that the mass of a gravitationally bound system, such as the solar system, a galaxy or a cluster of galaxies, is given by

$$M \approx D\langle v^2 \rangle / G \tag{14.36}$$

where D is the characteristic size of the system and $\langle v^2 \rangle$ a mean square velocity relative to the centre of mass. This is obviously true, for example, for a star of mass M with a small planet in a circular orbit of radius D with orbital velocity v. The masses of galaxies inferred in this way may be of the order of 10 times the mass that can be accounted for by visible stars. The *galactic halos* that contain this extra mass probably extend well beyond the visible part of the galaxy. For large clusters of galaxies, the inferred total mass may be several hundred times the mass of luminous matter. Since the critical density (14.24) is proportional to H^2, it might be thought that estimates of Ω_0 should depend on the value assumed for H_0. Actually, this is not so. The reason is that large distances are measured in terms of redshifts, by using the relation (14.11), and each distance measured in this way is thus proportional to H_0^{-1}. The *velocity dispersion* $\langle v^2 \rangle$ in (14.36) is estimated from distributions of redshifts around the mean for the object concerned and is independent of distance, so the estimate of mass is proportional to a distance estimate. Estimates of the density are therefore proportional to $(\text{distance})^{-2}$ or to H_0^2, and estimates of Ω_0 are independent of the value assumed for H_0. The value of Ω_0 which includes dark matter inferred from the virial theorem to exist in galactic halos and in the intergalactic medium in clusters is roughly in the range $0.1 < \Omega_0 < 0.3$.

If the cosmological constant is non-zero, then the deceleration parameter deduced from (14.20) with $p = 0$ is

$$q_0 = \tfrac{1}{2}\Omega_0 - \Omega_\Lambda \tag{14.37}$$

where $\Omega_\Lambda = \frac{1}{3}\Lambda H_0^{-2}$. There is a theoretical prejudice, arising from the *inflationary scenario* which I shall describe later, to the effect that the present universe ought to be almost exactly flat. In that case, Friedmann's equation with the addition of a cosmological constant as in (14.18) asserts that

$$\Omega_0 + \Omega_\Lambda = 1. \tag{14.38}$$

The observations of distant supernovae that I mentioned above seem to be consistent with this picture, if $\Omega_0 \approx 0.3$ and $\Omega_\Lambda \approx 0.7$ and this of course implies a negative value of q_0, which would mean that the expansion of the universe is currently accelerating. This would be a relatively recent phenomenon, however. It is easy to check that (14.23) is still valid when we allow for a non-zero value of Λ, and so is the conclusion that $\rho(t) \propto 1/a^3(t)$. In the two independent equations (14.18) and (14.20), therefore, Λ is insignificant compared with ρ at earlier times when $a(t)$ is much smaller than it is now. If Λ does turn out to be non-zero, then this will have an important bearing on the detailed fitting of cosmological models to observational data. However, the general picture will not be greatly affected, and I shall continue to simplify matters by taking $\Lambda = 0$.

From (14.13) and (14.34), our estimate of the present age of the universe is

$$t_0 = 9.78 \times 10^9 h^{-1} f(\Omega_0) \text{ years}. \tag{14.39}$$

If we take the limits on the observed parameters as $0.5 < h < 0.8$ and $0.1 < \Omega < 1$ then, referring to figure 14.4, we find that the age of the universe is within the limits

$$8 \times 10^9 \text{ years} < t_0 < 1.8 \times 10^{10} \text{ years}. \tag{14.40}$$

It is possible, of course, to place lower bounds on the age of the universe by estimating the age of objects it contains. Radio dating of terrestrial, lunar and meteoric rocks puts the age of the oldest material at about 4.5×10^9 years. It is believed, however, that this material is not primordial, but was formed in the cores of ancient stars, so the universe should be rather older than this. Estimates of the age of the oldest stars in our galaxy, those in globular clusters, suggest ages of about 10^{10} years. It is noteworthy that these independent estimates are in reasonable accord with those based on the cosmological parameters H_0 and Ω_0.

In addition to matter, the universe contains radiation. The most important component from a cosmological point of view is the *cosmic microwave background* which has, to a very good approximation, a black-body spectrum corresponding to a temperature of 2.7 K. This radiation, first observed by A A Penzias and R W Wilson (1965), is found to be isotropic to about one part in 10^4, except for a dipolar anisotropy which can be attributed to the motion of the Earth relative to comoving coordinates. This microwave background provides a vital clue to the early history of the universe. Because of its high degree of isotropy, it cannot have originated in observed galaxies, and is generally held to be a relic of an early period in which the content of the universe was a hot, dense plasma of particles and radiation. Because the universe must have been expanding very rapidly during this early phase, the standard cosmological model is often referred to as the *hot big bang* model. The importance of the microwave background for our present discussion is twofold. First, its isotropy provides the best evidence for the isotropy of the observable universe, on which the Robertson–Walker metric depends. Second, black-body radiation exerts a pressure as well as contributing to the energy density, so we can use it to estimate the duration of the matter-dominated era, during which the approximation $p = 0$ holds good.

From (10.91) with $g = 2$ for photons, we find for the energy density ρ_{rad} or the equivalent mass density ρ_{rad}/c^2 of the microwave background

$$\rho_{\text{rad}}(t_0) = 4.02 \times 10^{-14}\,\text{J m}^{-3} \quad \text{or} \quad \rho_{\text{rad}}/c^2 = 4.47 \times 10^{-34}\,\text{g cm}^{-3}. \qquad (14.41)$$

The overall density is given by (14.25) as

$$\rho(t_0) = \rho_{\text{matt}}(t_0) + \rho_{\text{rad}}(t_0) = h^2\Omega_0 \times 1.88 \times 10^{-29}\,\text{g cm}^{-3} \qquad (14.42)$$

which we can take as about $1 \times 10^{-29}\,\text{g cm}^{-3}$. At the present time, therefore, the contribution of the radiation to the energy density is negligible and its pressure, which is $\frac{1}{3}$ of its energy density, is also negligible in (14.22). However, we saw in (14.16) that the frequency or energy of a photon is proportional to $1/a(t)$. Therefore, if we assume a constant number of photons in a given comoving region, the energy density is proportional to $1/a^4(t)$, whereas that of non-relativistic matter is proportional to $1/a^3(t)$ as in (14.29). Thus we have $\rho_{\text{rad}}(t)/\rho_{\text{matt}}(t) \propto 1/a(t)$ and the radiation becomes more important at earlier times. To see how long the universe has been matter dominated, we can estimate the time t_{m} at which the densities of matter and radiation are approximately equal, which is also the time at which the radiation pressure becomes significant in

(14.22). The condition is

$$\frac{a(t_{\mathrm{m}})}{a(t_0)} = \frac{\rho_{\mathrm{rad}}(t_0)}{\rho_{\mathrm{matt}}(t_0)}\,\frac{\rho_{\mathrm{matt}}(t_{\mathrm{m}})}{\rho_{\mathrm{rad}}(t_{\mathrm{m}})} = \frac{\rho_{\mathrm{rad}}(t_0)}{\rho_{\mathrm{matt}}(t_0)} \cdot 1 \approx 5 \times 10^{-5}. \tag{14.43}$$

Since Ω is fairly close to 1, it is sufficient to use (14.31) to get

$$t_{\mathrm{m}}/t_0 \approx (5 \times 10^{-5})^{3/2} \approx 4 \times 10^{-7}. \tag{14.44}$$

This result shows that the time which has elapsed since the universe became matter dominated is about one million times that which had elapsed previously. For various reasons, it is not a very accurate estimate. Quite apart from the uncertainties in h and Ω_0, the relation $a(t) \propto t^{2/3}$ is not valid before t_{m}. A better approximation is, as we shall see, $a(t) \propto t^{1/2}$. This does not, however, invalidate the conclusion that t_{m} is only a tiny fraction of t_0 and, therefore, that the zero-pressure model can be used to estimate t_0 to the accuracy permitted by other uncertainties.

The foregoing argument assumes that the kinetic energy and pressure of matter, which is negligible now, was also negligible at t_{m}, so it would be as well to check that this is so. Since the energy density and pressure of the radiation are proportional to $1/a^4(t)$ and also, according to (10.93), to T^4, it follows that the temperature T is proportional to $1/a(t)$, in agreement with the conclusion we reached earlier by considering the energy of an individual photon. At $t = t_{\mathrm{m}}$, then, the temperature of the radiation was about 5×10^4 K. At this temperature, the matter consisted (as it turns out) mainly of ionized hydrogen and helium. This ionized matter interacts strongly with radiation and would have been in equilibrium with it at the same temperature. The equivalent energy $k_{\mathrm{B}}T$ is about 4 eV, so the kinetic energy even of the electrons was much smaller than their rest energy of 511 keV. Thus, the kinetic energy density and pressure of the matter was negligible compared with the density of its rest energy.

14.4 The Fairly Early Universe

Processes occurring in the early universe at temperatures below about 10^{12} K (at which $k_{\mathrm{B}}T$ is approximately equal to the mass of a muon, 106 MeV) can be investigated quite thoroughly on the basis of well established physics. This temperature probably occurred when the universe was about 10^{-4} s old, and I shall refer to the period between then and t_{m} when the universe became matter dominated as the 'fairly early' universe. The fairly early history of the universe has been carefully documented by, for example, Peebles (1971, 1993) and Weinberg (1972) and I shall largely follow Weinberg's account.

A few gross features are easily deduced. First of all, the curvature of space was unimportant. If space is curved, then $|k| = 1$ in (14.21). Let us define the ratio of k to the right-hand side of this equation as

$$K(t) = 3k/\kappa\rho(t)a^2(t). \tag{14.45}$$

From (14.24)–(14.26), we find that $K = (\Omega - 1)/\Omega$, and the limits on the present value of Ω imply for the present value K_0 of $K(t)$ that $|K_0| < 10$. During the matter-dominated era, in which $\rho(t) \approx \rho_{\mathrm{matt}}(t) \propto 1/a^3(t)$, we see that $K(t)$ is proportional to $a(t)$. Therefore, according to (14.43), we have $|K(t_{\mathrm{m}})| < 5 \times 10^{-4}$. At earlier times, $K(t)$ was even smaller, so k has a negligible effect in (14.21) and may be taken as zero. For much of the time, the early universe was *radiation dominated*, in the sense that its contents behaved like black-body radiation with $p = \frac{1}{3}\rho$, which is true not only for photons, but for any particles whose kinetic energy is much bigger than their rest energy. When this is true, (14.21) and (14.22) can be combined (with $k = 0$) to give

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(a^2\right) = 2\left(a\ddot{a} + \dot{a}^2\right) = 0 \tag{14.46}$$

whose solution is of the form

$$a(t) = (At + B)^{1/2}. \tag{14.47}$$

(Note also that setting $p = \frac{1}{3}\rho$ in (14.23) yields the relation between $\rho(t)$ and $a(t)$ analogous to (14.29) for a radiation dominated universe, namely $\rho(t) = \Gamma/a^4(t)$, where Γ is a constant, which we found earlier by a different argument.)

The assumption is usually made that ρ can be evaluated as the sum of densities of several species of particles, each behaving as an ideal gas in thermal equilibrium, and p as the sum of their pressures. Some thermodynamic relations that apply to these gases will be useful to us. Consider a comoving region, whose volume V is equal to a^3, and suppose that the particles it contains can be divided into groups such that the particles in each group interact with each other, but not with those in other groups. The idea is that, within each group, the interactions are strong enough for the temperature and relative numbers of particles to be determined by the condition of thermal equilibrium, but sufficiently weak that the interaction energy does not contribute significantly to the energy density and pressure. According to the fundamental relation (10.30) of equilibrium thermodynamics, we have for the ith group of particles

$$\frac{\mathrm{d}U_i}{\mathrm{d}t} + p_i\frac{\mathrm{d}V}{\mathrm{d}t} = T_i\frac{\mathrm{d}S_i}{\mathrm{d}t} + \sum_j^{(i)} \mu_j \frac{\mathrm{d}N_j}{\mathrm{d}t} \tag{14.48}$$

where the sum is over particle species belonging to the ith group. If we sum this over all groups of particles and use (14.23) with $\rho a^3 = \sum_i U_i$ and $p = \sum_i p_i$, we find

$$\sum_i T_i\frac{\mathrm{d}S_i}{\mathrm{d}t} = -\sum_j \mu_j\frac{\mathrm{d}N_j}{\mathrm{d}t} \tag{14.49}$$

where the sum on i is over all groups of mutually interacting particles and the sum on j is over all particle species.

It will sometimes be important to know the numbers of particles present as well as their contributions to the energy density and pressure. Using the basic distribution functions (10.64) and denoting by q the magnitude of a particle's 3-momentum, we find that the number of particles of a given species per unit physical volume with momentum in the range q to $q + \mathrm{d}q$ is

$$n(q)\mathrm{d}q = \frac{g}{2\pi^2\hbar^3}\left\{\exp\left[\beta(\epsilon(q) - \mu)\right] \pm 1\right\}^{-1} q^2\mathrm{d}q \qquad (14.50)$$

where $\epsilon(q) = c(q^2 + m^2c^2)^{1/2}$ and $g = 2s + 1$ is the spin multiplicity factor. If $\mu = 0$ and the particles are highly relativistic, so that their mass can be neglected, then the total number per unit volume is

$$n = \frac{g}{2\pi^2}\left(\frac{k_\mathrm{B}T}{c\hbar}\right)^3 \int_0^\infty (\mathrm{e}^x \pm 1)^{-1}x^2\mathrm{d}x = \frac{\zeta(3)}{\pi^2}\begin{pmatrix}3/4\\1\end{pmatrix} g\left(\frac{k_\mathrm{B}T}{c\hbar}\right)^3 \qquad (14.51)$$

where the upper values refer to fermions, the lower ones to bosons and ζ is the Riemann zeta function, with $\zeta(3) = 1.202\ldots$. At present, the microwave background contains some 400 photons per cm^3. If we take the present density of matter as $10^{-29}\,\mathrm{g\,cm^{-3}}$ and assume that it is primarily composed of nucleons, each with a mass of about 1.7×10^{-24} g, then the ratio of the number of nucleons to the number of photons, conventionally denoted by η is

$$\eta = n_\mathrm{N}/n_\gamma \approx 10^{-8}. \qquad (14.52)$$

If, on the other hand, much of the dark matter is not composed of nucleons (or, in the conventional terminology, is not 'baryonic'), then η might be much smaller than this. In fact, the theory of 'nucleosynthesis', about which I shall have something to say in the next section, strongly indicates that $\eta \sim 5 \times 10^{-10}$, which means that most of the dark matter is non-baryonic. What this matter might consist of is the subject of intense, and frequently exotic speculation, but no clear answer is known. At any rate, η is small, and it has been constant for most of the history of the universe; only at fairly early times were processes possible which could cause significant changes in either the number of photons or the number of nucleons in any comoving region.

We can now work out what conditions must have been like at temperatures a little below 10^{12} K. The nucleons that still exist today were present but their numbers, energy density and pressure were negligible compared with those of the black-body photons. The typical energy of a photon (a little less than 100 MeV) was such that electron-positron pairs could be copiously produced in collisions. These pairs could also annihilate to produce photons. Under the assumption of thermal equilibrium, the balance between these processes leads to a distribution of energies for the electrons and positrons of the form (14.50), and I shall shortly discuss the conditions under which this assumption is likely to be valid. Likewise, the electron- and muon-type neutrinos and antineutrinos could be produced and annihilated by weak interaction processes and also had a thermal distribution. The particles present in substantial numbers were therefore γ, e^-, ν_e and ν_μ, together with the antileptons, and there were also a few nucleons. All known heavier particles, which will have been present at higher temperatures, undergo rapid decays or particle-antiparticle annihilations, whose final products are the ones I have listed, and there was insufficient energy available to replenish them.

Under conditions of thermal equilibrium, the abundant particles have energy distributions of the form (14.50), so we need to know their chemical potentials. As we saw in chapter 10, the equilibrium density operator (10.59) can contain only operators associated with conserved quantities. For the particle species of interest, there are four conserved quantities, namely electric charge Q (measured in units of e), electron number E, muon number M and baryon number B. The values of these numbers for the various particles are

	e	ν_e	ν_μ	p	n
Q	-1	0	0	1	0
E	1	1	0	0	0
M	0	0	1	0	0
B	0	0	0	1	1

(14.53)

with opposite values for their antiparticles. These conservation laws are embodied in the standard GWS model. For example, any interaction vertex that creates an electron also either creates a positron or an anti-electron neutrino or annihilates an electron neutrino, so electron number is conserved. In grand unified theories, which allow processes like proton decay, the lepton and baryon numbers are not separately conserved. However, processes which violate these conservation laws will occur at significant rates only when collision energies are greater than the X boson masses of about 10^{15} GeV or at temperatures above 10^{27} K. In the density operator (10.59), we can introduce an independent chemical potential for each conserved quantity and then, using the values in (14.53), express Q, E, B and M in terms of particle numbers:

$$\begin{aligned}
&\mu_Q\hat{Q} + \mu_E\hat{E} + \mu_M\hat{M} + \mu_B\hat{B} \\
&\quad = \mu_Q\left[\hat{N}_{\mathrm{e}^+} - \hat{N}_{\mathrm{e}^-} + \hat{N}_\mathrm{p}\right] + \mu_E\left[\hat{N}_{\mathrm{e}^-} + \hat{N}_{\nu_\mathrm{e}} - \hat{N}_{\mathrm{e}^+} - \hat{N}_{\bar{\nu}_\mathrm{e}}\right] \\
&\qquad + \mu_M\left[\hat{N}_{\nu_\mu} - \hat{N}_{\bar{\nu}_\mu}\right] + \mu_B\left[\hat{N}_\mathrm{p} + \hat{N}_\mathrm{n}\right].
\end{aligned} \tag{14.54}$$

From this, we can read off the chemical potentials for the particle species themselves, For example,

$$\mu_{\mathrm{e}^+} = \mu_Q - \mu_E = -\mu_{\mathrm{e}^-}. \tag{14.55}$$

As in (10.66), we now adjust the chemical potentials to accommodate what we know or can guess about the mean numbers of particles. Consider the total electric charge. All the evidence is that this is now exactly zero so, since charge is conserved, it must have been zero in the early universe too. Adding up the charges of all the particle species, we have

$$Q = N_{e^+} - N_{e^-} + N_p = N(\mu_{e^+}) - N(\mu_{e^-}) + N_p = 0 \qquad (14.56)$$

where $N(\mu)$ is the integral of (14.50) with the electron mass and the appropriate chemical potential. Under the conditions we are considering, the numbers of electrons and positrons are comparable with the number of photons and thus, according to (14.52), very much greater than the number of protons. To a good approximation, therefore, the numbers of electrons and positrons must be equal. Thus, their chemical potentials must also be equal and, in view of (14.55), must vanish. I shall follow the usual assumption that the chemical potentials of the neutrinos also vanish, though there is no firm evidence for it. As in (14.55), we find that the chemical potential of a neutrino and its antiparticle are equal and opposite. Large neutrino chemical potentials lead to a condition called *degeneracy*, similar to that which characterizes electrons in metals. The consequences of neutrino degeneracy can be investigated, and the main effect is to increase the contribution of neutrinos to the total energy density. This in turn affects several predictions of the standard cosmological model, notably those for nucleosynthesis, which I discuss below. These effects serve to place constraints on the size of the chemical potentials, and interested readers will find some discussion of them in Weinberg (1972).

To continue the story of the fairly early universe, it is necessary to understand the conditions under which thermal equilibrium can be maintained. Readers will recall from chapter 10 that the ensemble averages of statistical mechanics correspond to long time averages for a single system. In order for the scattering processes that maintain the balance of particle numbers to be effective, it must be possible for a reasonable number of these events to occur before any great change takes place in the environment. To obtain a criterion for this, consider the mean free path λ of a particle between scattering events. Under laboratory conditions, it is given by $\lambda = 1/n\sigma$, where n is the number of particles per unit volume and σ is the scattering cross-section. In the expanding universe, consider a particle with velocity v relative to comoving coordinates, attempting to collide with a comoving target particle a distance λ away. The expansion is carrying the target particle away with a speed given by Hubble's law as $H\lambda$. A rough criterion for scattering to take place at a reasonable rate is that v should be considerably greater than $H\lambda$. Another way of putting this is that the mean time between collisions under laboratory conditions, $\tau = \lambda/v$ should be much less than the characteristic expansion time H^{-1}:

$$\tau H = H/n\sigma v \ll 1. \qquad (14.57)$$

Let us apply this to the weak interactions which are supposed to maintain the thermal distribution of neutrinos. The energies we are considering are much smaller than the masses of the weak gauge bosons, so the Fermi theory (with the addition of neutral currents) is adequate. Scattering cross-sections are proportional to G_F^2 where, as we saw in chapter 12, $G_F/(\hbar c)^3 = 1.17 \times 10^{-5}$ GeV^{-2}. Since $k_B T$ is much greater than the electron rest energy, it is the only relevant quantity with the dimensions of energy, and dimensional analysis shows that the cross-sections must be given by

$$\sigma \approx G_F^2 (k_B T)^2 (\hbar c)^{-4}. \tag{14.58}$$

If we take $H = (\kappa\rho/3)^{1/2}$, ρ and n to be given by the thermal distributions for a few species of particles and, for neutrinos and highly relativistic electrons, $v = c$, we obtain the estimate

$$\tau H \approx (10^{10}/T)^3 \tag{14.59}$$

when T is measured in degrees Kelvin. As our story starts, just below 10^{12} K, this is small enough for thermal equilibrium to become established, if it had not already been, and to be maintained. As the temperature fell to around 10^{10} K, however, the rate of neutrino scattering became very small so that, in effect, the neutrinos ceased to interact with the other particles or, as the jargon has it, became *decoupled*. The thermal distributions of neutrinos did not disappear, however. Their temperature simply continued to fall as $1/a(t)$ and they are, presumably, here to this day, though it would be extremely difficult to detect them. Their present temperature is, as we are about to see, rather less than that of the microwave background and their contribution to the energy density correspondingly smaller. The cross-section for electromagnetic scattering of electrons, positrons and photons is greater than the weak cross-sections, and these particles continued to interact.

The rest energy of an electron corresponds to a temperature of about 5.9×10^9 K. As the temperature dropped below this value, electron-positron pairs could no longer be produced by collisions. The electrons and positrons which had been present annihilated rapidly, producing extra photons which heated the black-body radiation. Since the neutrinos had ceased to interact, their temperature was unaffected, so the temperature of the photons was now greater than that of the neutrinos and has remained so ever since. We can work out the ratio of the photon and neutrino temperatures from (14.49). The right-hand side of this equation is zero, as may be seen in the following way. The chemical potentials of the electrons, positrons and photons are zero. The only other particles present in significant numbers are the neutrinos and, since these have ceased to interact, the number of them in a comoving volume is constant. So, regardless of their chemical potentials, neutrinos do not contribute to the right-hand side of (14.49). On integrating (14.50) for a neutrino species, with $m = 0$, we find that the total number in a comoving volume proportional to a^3 can be expressed as $(aT)^3 f(\mu/T)$, where f is the function determined by the integral and multiplying constants. Since this number is constant, and T is proportional to $1/a$, the ratio μ/T is constant. The neutrino entropy in the comoving volume can, as readers may easily check, be expressed in the same form, but with a different function f, so it too is constant and makes no contribution to the left-hand side of (14.49). Thus, the left-hand side of (14.49) has significant contributions only from electrons, positrons and photons which, since they still interact, have the same temperature. What (14.49) tells us, therefore, is that the total entropy of electrons, positrons and photons in a comoving volume is constant, regardless of the neutrino chemical potentials.

While the electron-positron annihilation is taking place, the electron mass is comparable with $k_{\mathrm{B}}T$, and the integral for the entropy cannot be computed analytically. For our present purpose, however, this is not necessary. We consider a time 'before' the annihilation when the electrons were relativistic, and a time 'after' the annihilation when they had vanished. In each case, we can use (10.92) for the electron-positron-photon entropy. The multiplicity factor g is given by (10.88) as

$$g_{\text{before}} = 2 + \tfrac{7}{8} \times 4 = \tfrac{11}{2} \qquad \text{and} \qquad g_{\text{after}} = 2 \tag{14.60}$$

since the electron, positron and photon each have two polarizations. Conservation of this entropy implies

$$g_{\text{before}}(aT)^3_{\text{before}} = g_{\text{after}}(aT)^3_{\text{after}} \tag{14.61}$$

where T is the photon temperature. For the neutrino temperature T_ν, on the other hand, we have

$$(aT_\nu)_{\text{after}} = (aT_\nu)_{\text{before}} = (aT)_{\text{before}} \tag{14.62}$$

and so, after the annihilation

$$T_\nu = (g_{\text{after}}/g_{\text{before}})^{1/3}T = (4/11)^{1/3}T = 0.714\,T. \tag{14.63}$$

The present neutrino temperature is therefore about 1.9 K.

As far as the abundant species of particles are concerned, there were no further significant events until the universe became matter dominated. The state of the nucleons did indeed undergo important changes, which are discussed in the next section, but these had no significant effect upon the energy density, pressure or expansion rate. We can now estimate the periods of time that elapsed between the various events I have described so far. Consider a period during which the multiplicity factor g^* for the total number of abundant species is constant. (Note that g^* is different from the g given in (14.60), which counts only those particles interacting efficiently with photons.) Since we have set $k = 0$, we may use equations (14.47), (14.10), (14.24) and (10.91) to express a time difference $t_2 - t_1$ in terms of the temperatures T_2 and T_1 prevailing at those times. The result is

$$\begin{aligned} t_2 - t_1 &= \left(\frac{3c^2}{64\pi G\sigma}\right)^{1/2} g^{*-1/2}\left(T_2^{-2} - T_1^{-2}\right) \\ &= 3.26 \times 10^{20} g^{*-1/2}\left(T_2^{-2} - T_1^{-2}\right) \end{aligned} \tag{14.64}$$

where the times are in seconds and temperatures in degrees Kelvin.

In order to make use of this result, we need to know the value of g^*, which means that we need to know all the species of particles which were present. We have seen that the electron- and muon-type neutrinos were decoupled at temperatures below about 10^{10} K but still contributed to the energy density and pressure. However, we also saw in chapter 12 that a further neutrino species, the tau-type neutrino, is known to exist. These and perhaps other, as yet unknown, species of neutrinos or other light particles will also have been present. Whatever these species are, we know from laboratory experiments that they do not interact strongly at the temperatures we have considered, so they do not affect our calculations up to this point. They will, however, affect any calculations that require us to know periods of time rather than merely temperatures, and this is one point at which theoretical models of particle physics have cosmological consequences which can be confronted with observations. Each additional species has, presumably, a thermal energy distribution similar to that of the neutrinos. As we have seen, however, the temperature of the electron and muon neutrinos was changed relative to that of the photons by the electron-positron annihilation. Depending on the temperature at which a given species decoupled, its temperature may have been similarly affected by earlier annihilation processes, of which we have no definite understanding.

These matters can be dealt with in detail only on the basis of some definite model of particle physics and, in general, some additional assumptions about the sequence of events in the very early universe, at temperatures above 10^{12} K. For the sake of argument, I shall suppose that there are N_ν species of neutrinos, all at the same temperature. In that case, the value of g^* prior to electron-positron annihilation is

$$g^* = \tfrac{11}{2} + \tfrac{7}{4}N_\nu \qquad \text{for } 10^{12}\text{ K} > T > 6 \times 10^9\text{ K} \tag{14.65}$$

assuming that each neutrino and its antiparticle together contribute two polarization states. After the annihilation, we can take account of the different neutrino temperature by including an appropriate factor in g^*:

$$g^* = 2 + \tfrac{7}{4}\left(\tfrac{4}{11}\right)^{4/3} N_\nu \qquad \text{for } T < 6 \times 10^9\text{ K}. \tag{14.66}$$

Let us calculate some representative time intervals, taking $N_\nu = 3$ to include just the three known neutrinos. The time taken for the temperature to fall from 10^{12} K to 10^{11} K was 9.8×10^{-3} s. The further time to reach 10^{10} K was, obviously, a hundred times this, 0.98 s. Near their annihilation temperature, the electrons and positrons are non-relativistic, so our equations based on black-body radiation are not valid, and a numerical calculation using the correct distribution is needed. It is a fair approximation, however, to imagine that the annihilation occurred instantaneously, using (14.65) just above and (14.66) just below 6×10^9 K. With this approximation, the time to get from 10^{10} K to 6×10^9 K was 1.77 s, and the further time to reach 10^9 K was 4.9 hours. According to (14.43), the universe became matter dominated at a temperature of $2.7/5 \times 10^{-5} = 5.4 \times 10^4$ K. If we use (14.64) to estimate when this happened, the answer is about 2,000 years after the events we have been considering. This calculation is not quite right, though, because the basic equation (14.64) assumes that the density of matter is negligible. Also, our estimate of the temperature at which the densities of matter and radiation became equal depends on the values assumed for H_0 and Ω_0. The actual time is therefore not very accurately known, but it is most often estimated at about 10,000 years.

To estimate the time from the initial singularity to our starting point at 10^{12} K, we would need to know what happened during that time. If we assume that (14.64) remains valid, then the value of g^* obviously increases with temperature. Thus it is reasonable to guess that this time is no greater that what we obtain by using (14.65) and setting the initial temperature to infinity, namely about 10^{-4} s. Clearly, using the figures given above, we might as well say that the temperature was 10^{10} K at 0.98 s after the initial singularity, and so on.

14.5 Nucleosynthesis

Although protons and neutrons made a negligible contribution to the overall composition of matter in the early universe, they were nevertheless able to take part in interactions which had important consequences. There is a narrow range of temperatures around 10^9 K at which nuclear reactions could take place which fused protons and neutrons into larger nuclei. These reactions have been well studied in the laboratory, and it is possible to work out quite accurately the relative numbers in which various light nuclei would have been formed. The process is called *nucleosynthesis* and it is important for at least two reasons. On the one hand, the predicted abundances can be compared with matter actually observed in the present universe, and after allowance has been made for later reactions occurring in the cores of stars, the overall agreement is found to be rather good. This provides an important test of the standard big bang model. On the other hand, the predicted abundances of some nuclei depend on the values of quantities such as N_ν in (14.65) and the density of nucleons available to form nuclei. The comparison with observations then serves to determine the values of these quantities, or at least to put useful constraints on their possible values. It turns out that hydrogen and helium-4 are by far the most abundant nuclear species, and I shall give a simplified account of the calculation of their relative abundances. Interested readers will find more details and further references in, for example, Peebles (1971), Weinberg (1972), Barrow (1983) and Bernstein *et al* (1989); a survey of recent developments is given by Schramm and Turner (1998).

The relative abundances of nuclei obviously depend on the relative numbers of protons and neutrons and, to estimate their ratio, we must begin the story of nucleosynthesis at a temperature of about 10^{11} K. Although the total number of nucleons cannot change at this temperature (the typical energy $k_BT \sim 9$ MeV is much smaller than the nucleon rest energy of about 940 MeV), lepton-nucleon scattering can easily interconvert protons and neutrons by weak-interaction processes such as $\mathrm{e}^- + \mathrm{p} \leftrightarrow \mathrm{n} + \nu_\mathrm{e}$. The energy absorbed or released by these conversions is the neutron-proton mass difference $\Delta m = m_\mathrm{n} - m_\mathrm{p} = 1.29$ MeV. As long as the weak interactions are effective in maintaining thermal equilibrium, the ratio of the numbers of protons and neutrons can adequately be determined from classical statistical mechanics and is given by

$$n_\mathrm{n}/n_\mathrm{p} = \exp\left(-\Delta m/k_\mathrm{B}T\right). \tag{14.67}$$

At about the time the neutrinos cease to interact with electrons, the interconversions of protons and neutrons also cease, and the ratio becomes frozen.

For good accuracy, it is necessary to determine the ratio precisely, and this requires a detailed analysis of the reaction rates, which I am not going to reproduce here. It is easy to see, however, that the ratio depends on the value of g^* at the temperature T_f where the freeze occurs. Consider, for example, neutron-neutrino scattering, for which the cross-section is roughly the same as (14.58). As readers may convince themselves, the number of scattering events per unit time per unit volume is $\sigma n_\nu n_\mathrm{n} c$, where n_ν and n_n are the number densities of neutrinos and neutrons respectively. The number of events per unit time per neutron is therefore $\sigma n_\nu c$. The mean time between scattering events for a particular neutron is $1/(\sigma n_\nu c)$ and, roughly speaking, the freeze occurs when this time equals the expansion time H^{-1}. To estimate T_f, we use (14.51) with $g = 1$ for n_ν, (14.24) for H and estimate ρ using (10.91) with g equal to the g^* given in (14.65) for all the abundant species present at temperatures near 10^{10} K. The result is

$$T_\mathrm{f} \approx 2.6 \times 10^{10} g^{*1/6}\,\mathrm{K}. \tag{14.68}$$

Inserting this value into (14.67) gives a good indication of how the neutron-proton ratio depends on g^* and hence on N_ν, but the number 2.6×10^{10} is merely a guess. The results of a more careful analysis, insofar as they can be approximated by an equation of the form (14.67), indicate that this number should be replaced by something like 6.4×10^9.

At the prevailing nucleon densities, the probability of more than two particles colliding simultaneously is negligible, so nuclei can be built up only by two-particle collisions. The first nucleus that can be formed is deuterium, consisting of one proton and one neutron. Now, deuterium has a binding energy of only about 2.2 MeV and, at temperatures near 10^{10} K, there are many photons capable of dissociating it. Deuterium nuclei remain intact in sufficient numbers for further reactions to proceed only when the temperature has fallen to a value which is estimated at about 8×10^8 K. This value depends somewhat on the actual numbers of nuclei present, which in turn are related to the present matter density. Studies of the reactions which then ensue show that almost all of the available neutrons are used to form helium-4, the excess protons remaining single. Only very small quantities of heavier nuclei such as lithium-7 emerge, together with small amounts of deuterium and tritium.

The relative abundance of hydrogen (protons) and helium is thus essentially determined by the neutron-proton ratio at 8 × 10^8 K, and I shall now estimate it, taking N_ν to be 3. At the temperature T_f, which is 9.5 × 10^9 K, the ratio $n_\mathrm{n}/n_\mathrm{p}$ is given by (14.67) to be 0.206, and the fraction $X_\mathrm{n} = n_\mathrm{n}/(n_\mathrm{p} + n_\mathrm{n})$ is 0.171. The time that elapses as the temperature falls from T_f to 8 × 10^8 K is found from (14.64) to be 274 s. During this time, a few neutrons decay, each one to a proton plus leptons, with a mean lifetime of 917 s, so when nucleosynthesis begins we have

$$X_\mathrm{n} = \frac{n_\mathrm{n}}{n_\mathrm{p} + n_\mathrm{n}} = 0.171 \exp(-274/917) = 0.127. \tag{14.69}$$

Since each ^{4}He nucleus contains two neutrons and two protons and has almost exactly four times the mass of a proton, the fraction by weight of helium, $M_\mathrm{He}/(M_\mathrm{He} + M_\mathrm{H})$, is, as readers may check, just twice this number, or about 25.3%. I emphasize that, while this calculation illustrates the essential argument, a much more thorough analysis is needed to obtain reliable results. A detailed analysis does predict a ^{4}He abundance of around 25% (the value favoured by Schramm and Turner (1998) is 0.248 ± 0.002), but the values obtained for the abundances of light nuclei depend both on N_ν, which affects the expansion rate, and on the density of nucleons (or baryons), ρ_B.

The abundance of ^{4}He turns out to depend very little on ρ_B, but it does depend significantly on N_ν. In my schematic calculation, the dependence on N_ν is through g^*. Taking $N_\nu = 2$ would give a ^{4}He abundance of 23.9%, while $N_\nu = 4$ would give 26.7% and detailed calculations give variations of about the same size. In the 1980s, comparison of these results with observational estimates of the primordial abundance of ^{4}He made it possible to place an upper limit of 4 on the number of species of neutrinos—a number which from direct particle-physics considerations was known only to be smaller than 8. [According to the standard model of particle physics, N_ν is the same as the number of families of fermions, and the number of quark flavours is $2N_\nu$. As we saw in connection with (12.63), asymptotic freedom, which seems to be a well-verified property of QCD, is valid only if there are no more than 16 quark flavours, which implies $N_\nu \leq 8$.] In 1989, a direct determination of N_ν became possible through measurement of the lifetime (or, more accurately, the decay width) of the weak vector boson Z^0, which can decay into $\nu\bar{\nu}$ pairs of any species (see Abe *et al* (1989), Abrams *et al* (1989), Adeva *et al* (1989), Decamp *et al* (1989)). From measurements of this kind, N_ν is now known to be equal to 3 with negligible error.

By contrast, the abundance of deuterium, which is of the order of 3 × 10^{-5}, depends strongly on the baryon density ρ_B. Recent, accurate determinations of the deuterium abundance (Burles and Tytler (1998)) indicate that, if nucleosynthesis calculations are correct, then the contribution of ordinary nuclear matter to the present total density is $\Omega_\mathrm{B} h^2 \simeq 0.02$. According to our discussion in §14.3, direct estimates of the total matter density are generally consistent with a value of $\Omega_0 \sim 0.3$, so the existence of large quantities of some kind of non-baryonic dark matter seems to be strongly indicated.

14.6 Recombination and the Horizon Problem

By the time of nucleosynthesis, almost all the electrons and positrons that had once been present had annihilated. Assuming, however, that the universe is electrically neutral, there must have been a small residual number of electrons to balance the charge of the protons. When the temperature fell to a small enough value, T_r, these electrons will have combined with the positive nuclei to form neutral atoms. To estimate T_r with reasonable accuracy, it is sufficient to consider a universe filled entirely with hydrogen. Near T_r, the fraction x of ionized atoms is determined by thermal equilibrium, maintained by atomic collisions, and this is described by the *Saha equation* (exercise 10.9). This equation involves the number density of protons, which can be expressed in terms of the density of photons and the nucleon-photon ratio η. Taking the ionization energy as 13.6 eV, we obtain

$$x^2/(1-x) = 1.19 \times 10^{14} \eta^{-1} T^{-3/2} \exp(-1.578 \times 10^5/T). \tag{14.70}$$

A numerical solution of this equation is easy. If we take $\eta \approx 10^{-10}$, as is implied by the prediction $\Omega_\mathrm{B} h^2 \simeq 0.02$ of nucleosynthesis, then we find that x falls quite swiftly from a value close to 1 at $T = 4,000$ K to a very small value at $T = 3,000$ K.

While electromagnetic radiation interacts strongly with charged particles, it interacts hardly at all with a gas of neutral hydrogen and helium, which is almost completely transparent. It follows that the microwave background we observe today was last scattered at the time of recombination and has travelled freely towards us ever since. This leads to a conundrum known as the *horizon problem*, which I shall now explain. The path of a light ray is found by setting $\mathrm{d}\tau = 0$ in (14.1) where, for simplicity, I shall take $k = 0$. As measured by comoving coordinates, the distance it travels between times t_1 and t_2 is

$$L = \int_{t_1}^{t_2} \frac{\mathrm{d}t}{a(t)}. \tag{14.71}$$

Recombination occurred, as readers may work out, somewhat after the universe became matter dominated. For simplicity again, however, I shall assume that $a(t)$ was proportional to $t^{1/2}$ right up to recombination, since this will not greatly affect our conclusion. The coordinate distance d which a non-interacting light ray could have travelled between the initial singularity and the time t_r of recombination is

$$d = 2t_\mathrm{r}/a(t_\mathrm{r}). \tag{14.72}$$

Of course, light rays did interact strongly. The point is that no signal of any kind could have travelled a distance greater than d, and so any causal influences could have acted only within a 'causally connected' region whose diameter was no greater than d, which is called the *causal horizon*.

Since recombination, the universe has been matter dominated and, to a reasonable approximation, we can use the scale factor (14.31) to write

$$\frac{a(t)}{a(t_\mathrm{r})} = \left(\frac{t}{t_\mathrm{r}}\right)^{2/3}. \tag{14.73}$$

Then the coordinate distance D which a photon we now detect has travelled towards us since recombination is

$$D = \frac{3t_\mathrm{r}}{a(t_\mathrm{r})}\left[\left(\frac{t_0}{t_\mathrm{r}}\right)^{1/3} - 1\right] \simeq \frac{3t_\mathrm{r}}{a(t_\mathrm{r})}\left[\frac{a(t_0)}{a(t_\mathrm{r})}\right]^{1/2}. \tag{14.74}$$

The angle subtended at the Earth by one causally connected region is the ratio

$$\frac{d}{D} = \frac{2}{3}\left[\frac{a(t_\mathrm{r})}{a(t_0)}\right]^{1/2} = \frac{2}{3}\left(\frac{T_0}{T_\mathrm{r}}\right)^{1/2} \approx 0.02\,\mathrm{rad} \approx 1^\circ. \tag{14.75}$$

What is puzzling about this is that the observed radiation is completely isotropic. Thus, at the time of recombination, very many regions which could never have communicated with each other were, to at least one part in 10^4, at the same temperature.

14.7 The Flatness Problem

Cosmologists speak of a second puzzle concerning the standard model, which is called the *flatness problem*. During the whole history of the universe, the scale factor $a(t)$ has been roughly proportional to a power of t, say t^x with x equal to either $\frac{1}{2}$ or $\frac{2}{3}$. To make matters simple, suppose that x was always $\frac{1}{2}$. Crudely, we can then use (14.26) to compare the present density with that at earlier times:

$$\Omega(t) - 1 \approx \left(\frac{\dot a(t_0)}{\dot a(t)}\right)^2 (\Omega_0 - 1) \approx \left(\frac{t}{t_0}\right)(\Omega_0 - 1). \tag{14.76}$$

It will be recalled that the value $\Omega = 1$ corresponds to a flat universe, and it seems most unlikely that Ω_0 could differ from this value by more than a factor of 100. When the universe was, say, 1 second old, Ω must have been equal to 1 with an accuracy of at least one part in 10^{15}, and this seems to represent a degree of fine tuning which would not be expected to occur without some good reason.

Whether this should be regarded as a genuine puzzle is to some extent a matter of philosophical taste. Even though (14.76) is not exactly correct, it is obvious that, whatever the value of Ω_0, we shall find a value of $\Omega(t)$ that is arbitrarily close to 1 if we choose a sufficiently early time. It is worth reflecting, however, that all the events which determined the overall constitution of the universe took place within the first few seconds, if we are content to regard nucleosynthesis as a relatively minor rearrangement of the particles that already existed. Thus, all the relevant time scales that naturally arise from physics are of the order of a second or less and, unless Ω is for some reason exactly equal to 1, we might have expected some appreciable variation by that time. It is sometimes said, indeed, that the only truly fundamental time scale is the Planck time (14.28), at which $|\Omega - 1|$ was less than 10^{-60} or so, and that we might have expected some appreciable difference of Ω from 1 by then. At any rate, if Ω is exactly equal to 1, then we would certainly like to know why. If it is not, then, since $|\Omega - 1|$ grows with time at least as fast as $t^{1/2}$, we may reasonably wonder why the difference is still fairly small after some 10 billion years.

The horizon and flatness problems do not make the standard cosmological model incompatible with observations, but they do seem to show that the model requires very special initial conditions. Any explanation of these initial conditions must be sought in the very early universe, at temperatures well above 10^{12} K.

14.8 The Very Early Universe

As we attempt to look back into the very early universe, by which I mean the first 10^{-4} s, we soon encounter energies of a few hundred GeV at which the standard model of particle physics has been only incompletely tested in the laboratory. (Readers may like to bear in mind that an energy of 1 GeV corresponds to a temperature of 1.16×10^{13} K.) At still higher energies, the standard model may well be quite inadequate. It is widely thought that the grand unified and/or supersymmetric theories that we touched on in chapter 12 or the string theories to be discussed in chapter 15 should come into play, but there is no firm experimental foundation for any of these theories. Little of what is said about the very early universe can therefore be taken as reliably established and much of it is purely conjectural. As I said at the beginning of this chapter, however, it is possible in principle to work out some of the consequences of these theoretical conjectures and confront them with observations.

It seems that a prominent role must have been played by *phase transitions* of various kinds. The first of these that we encounter, moving backwards in time, is the *quark-hadron* or *deconfinement* transition. The idea is that, at sufficiently high temperature and density, quarks and gluons cease to be bound in identifiable hadronic particles, but exist instead in a relatively weakly interacting plasma along with the photons and leptons. Approximate calculations based on the lattice version of QCD suggest that this change takes place at a sharp phase transition which, at the fairly low density of nucleons present in the early universe, would have occurred at a temperature of around 10^{12}–10^{13} K. Experimental studies of heavy-ion collisions, which produce, for a short time, large densities of nuclear matter at high energy, provide some evidence for this kind of effect. Deconfinement is related to the property of *asymptotic freedom* which means, as readers will recall from chapter 12, that the effective strength of the strong interactions decreases at high energy. Were it not for asymptotic freedom, indeed, very little could be said at all about the first millisecond. Most of what we believe about the fairly early universe is based on treating radiation and matter as nearly ideal gases. If the 'strong' interactions continued to be strong at nucleon densities approaching those in atomic nuclei, then the difficulty of applying statistical mechanics to such a strongly interacting fluid would become prohibitive. If the idea of asymptotic freedom is correct, then we do not encounter such densities until the temperature is high enough, and the strong interaction weak enough, for the ideal gas approximation to be adequate.

If the gauge theories of fundamental interactions are correct, then we may expect phase transitions to occur at which their symmetries cease to be spontaneously broken. The possibility of symmetry restoration at high temperatures was first recognized by D A Kirzhnits and A D Linde (1972). These phase transitions are quite analogous to the superconducting transition, with critical temperatures given, very roughly, by the masses of the relevant gauge bosons.

To indicate how this works, I shall consider a single scalar field ϕ, which could be one of the Higgs fields in a gauge theory. For simplicity, I shall take it to be real, with a finite-temperature action similar to (10.76) given by

$$S_\beta(\phi) = \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \left[\frac{1}{2}\left(\frac{\partial\phi}{\partial\tau}\right)^2 + \frac{1}{2}\nabla\phi \cdot \nabla\phi + \frac{\lambda}{4!}(\phi^2 - v^2)^2 \right]. \quad (14.77)$$

Up to loop corrections in perturbation theory, the vacuum expectation value of ϕ is one of the two values $\pm v$, which are the two minima of the potential term in (14.77). A high-temperature state is, however, not a vacuum state, and we need to estimate the expectation value of ϕ in this state. To that end, we introduce a source J for the field and, as in (10.80), define a thermodynamic potential by

$$\begin{aligned} \exp[-\beta V \Omega(\beta, J)] &= Z_{\mathrm{gr}}(\beta, V, J) \\ &= \int \mathcal{D}\phi \, \exp\left[-S_\beta + J \int \mathrm{d}\tau \mathrm{d}^3x \, \phi(\boldsymbol{x}, \tau) \right]. \end{aligned} \quad (14.78)$$

The expectation value of ϕ should be independent of $\boldsymbol{x}$ and τ and is given by

$$\bar{\phi} \equiv \langle\phi\rangle_\beta = -\left.\frac{\partial\Omega}{\partial J}\right|_\beta. \quad (14.79)$$

Consequently, the thermodynamic relation analogous to (10.32) is

$$\mathrm{d}\Omega = -s\mathrm{d}T - \bar{\phi}\mathrm{d}J \quad (14.80)$$

where s is the entropy density. For the free energy $F(\beta, \bar{\phi})$ defined by the Legendre transformation

$$F(\beta, \bar{\phi}) = \Omega + J\bar{\phi} \quad (14.81)$$

we have

$$\mathrm{d}F = -s\mathrm{d}T + \bar{\phi}\mathrm{d}J \tag{14.82}$$

and consequently

$$\left.\frac{\partial F}{\partial \bar{\phi}}\right|_{\beta} = J. \tag{14.83}$$

Thus, when J is zero, the expectation value we require is a minimum of F which, as we shall see, is equal to the potential in (14.77) plus a temperature-dependent correction.

A satisfactory calculation of F is slightly complicated, but I shall present a simple calculation that captures the main result. The calculation is essentially first-order perturbation theory. We write ϕ as $\bar{\phi} + \psi$ and expand S_β to quadratic order in ψ, leaving out the interaction terms:

$$\begin{aligned} S_\beta(\phi) = \beta V \left[\frac{\lambda}{4!}(\bar{\phi}^2 - v^2)^2 - J\bar{\phi}\right] \\ + \int_0^\beta \mathrm{d}\tau \int \mathrm{d}^3x \left[\frac{1}{2}\left(\frac{\partial\psi}{\partial\tau}\right)^2 + \frac{1}{2}\nabla\psi\cdot\nabla\psi + \frac{1}{2}m^2(\bar{\phi})\psi^2\right] \end{aligned} \tag{14.84}$$

where

$$m^2(\bar{\phi}) = \frac{\lambda}{6}(3\bar{\phi}^2 - v^2). \tag{14.85}$$

To lowest order, the expectation value $\bar{\phi}$ is the value of ϕ that minimizes the quantity $S_\beta - J\int \mathrm{d}\tau\mathrm{d}^3x\,\phi$, so the term linear in ψ can be omitted. Next, we estimate Ω by substituting this into (14.78) and carrying out the functional integral, which is similar to the one which led to (10.84), except that we now have only one particle species. The result for the free energy (14.81) is

$$\begin{aligned} F(\beta,\bar{\phi}) = \frac{\lambda}{4!}\left(\bar{\phi}^2 - v^2\right)^2 \\ + \frac{1}{2\pi^2\beta^4}\int_0^\infty \mathrm{d}x\,x^2 \ln\left\{1 - \exp\left[-\left(x^2 + \beta^2 m^2(\bar{\phi})\right)^{1/2}\right]\right\}. \end{aligned} \tag{14.86}$$

(At higher orders, the term linear in ψ cannot be neglected. A more systematic procedure is to determine J as a function of $\bar{\phi}$ from the requirement that $\langle\psi\rangle_\beta = 0$ and use (14.83) to find F.)

The import of this result becomes clearer if we make a high-temperature expansion, whose first few terms are

$$F(\beta,\bar{\phi}) = \left[\frac{\lambda}{24}v^4 - \frac{\pi^2}{90}(k_\mathrm{B}T)^4\right] + \frac{\lambda}{12}\left[\frac{1}{4}(k_\mathrm{B}T)^2 - v^2\right]\bar{\phi}^2 + \frac{\lambda}{24}\bar{\phi}^4 + \ldots. \tag{14.87}$$

This is similar to a Ginzburg–Landau expansion. The coefficient of $\bar{\phi}^2$ can be thought of as a temperature-dependent effective mass for the ϕ particles, which characterizes the way in which they propagate through a plasma of other particles. We see that the critical temperature at which symmetry is restored is given by $k_{\mathrm{B}}T_{\mathrm{c}} = 2v$. When ϕ is a Higgs field, this critical temperature is related to gauge-boson masses at zero temperature by equations similar to (12.24), so unless the gauge coupling constant is very large or very small, these masses give a fair indication of T_{c}. In this approximation, the expectation value of ϕ is clearly given by $\bar{\phi} = \pm v\left[1 - (T/T_{\mathrm{c}})^2\right]^{1/2}$.

If a phase transition of this kind leads to restoration of the SU(3) × SU(2) × U(1) symmetry of the standard model of particle physics, then this occurred at a temperature around 10^{15} K, at a time of about 10^{-12} s. It does not appear that this would have had any great effect on the expansion rate. In the case of a grand unified theory, the transition would occur at a temperature of some 10^{27} K, about 10^{-35} s after the initial singularity. According to what is called the *inflationary scenario*, the effect of this may have been spectacular. The idea of inflation was proposed by A Guth (1981) as a possible solution to the horizon and flatness problems, and also as a means of explaining the absence from the known universe of magnetic monopoles which ought, so it would seem, to be produced at a GUT phase transition through the Kibble mechanism that I touched on in §13.3. According to Guth, the universe may, at a very early time, have undergone a short period of very much more rapid expansion than is envisaged in the standard cosmological model.

To see how this might come about, consider a period during which the temperature is falling towards the critical temperature for a symmetry-breaking phase transition, involving a scalar field with an action similar to (14.77). The expectation value of ϕ is zero, which at this point is the state of minimum free energy. Below T_{c}, the state of thermal equilibrium is one in which the expectation value is non-zero, but the field will require some period of time to adjust to this new state. During this time, equilibrium statistical mechanics is not valid. What we should use in its place is a difficult question to which no satisfactory answer has (in my view) been found, but an obvious starting point is to obtain the stress tensor for the field ϕ, which should appear in the field equations in place of the stress tensor of an ideal gas that we have used until now. A general expression, implicit in the derivation of the field equations of general relativity (see exercise 4.2) is

$$T^{\mu\nu} = -\frac{2}{\sqrt{-g}}\frac{\delta S}{\delta g_{\mu\nu}}. \tag{14.88}$$

For a real scalar field, with potential $V(\phi)$, the action in a curved spacetime (see §7.7) might be taken as

$$S = \int \mathrm{d}^4x\,(-g)^{1/2}\left[\tfrac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right] \tag{14.89}$$

if we assume a minimal coupling to the spacetime curvature, and the stress tensor is

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left[\tfrac{1}{2}\partial_\sigma\phi\partial^\sigma\phi - V(\phi)\right]. \tag{14.90}$$

At this point, we meet a serious difficulty of principle, because the field equations $G_{\mu\nu} = \kappa T_{\mu\nu}$ equate the components of the tensor $G_{\mu\nu}$, which describe the geometry of a *classical* spacetime manifold and have definite values at each point of the manifold, to those of a *quantum-mechanical operator* $T_{\mu\nu}$, which act in the Hilbert space of the field theory and have no definite values at all. In a consistently quantum-mechanical description of the world, we would presumably have some analogous equation relating operators associated with both the spacetime geometry and its matter content. Having no such theory in hand, however, we must find some stop-gap means of making sense of the incomplete theories that we do have. This difficulty has, of course, been lurking in the background ever since we started to use the ideas of particle physics to work out the behaviour of matter in the early universe. We have avoided it until now because equilibrium statistical mechanics appears to give us an unambiguous way of calculating the pressure and density of a fluid, whether we imagine the particles in the fluid to be classical or quantum-mechanical ones. The crucial step is contained in definitions such as (10.60), where the trace incorporates averages over both statistical uncertainties and quantum indeterminacy. In effect, the assumption underlying our cosmological considerations has been that the field equations can be taken as

$$G_{\mu\nu} = \kappa\langle T_{\mu\nu}\rangle \tag{14.91}$$

where the expectation value uses an equilibrium density operator of the kind given in (10.59). The generalization of these equations to encompass expectation values in a state that may not be one of thermal equilibrium may be called the *semi-classical Einstein equations*. While they cannot be justified at a fundamental level, the success of quantum statistical mechanics in dealing with both equilibrium and non-equilibrium situations in condensed matter physics offers encouragement that they may give us roughly the right answers in a cosmological setting also.

In the case of an isotropic universe, described by the Robertson–Walker metric, the quantum-mechanical state must respect the assumption of isotropy. In particular, this implies that the spatial components of the stress tensor obey $\langle T_{ij}\rangle = \frac{1}{3}\delta_{ij}\sum_k\langle T_{kk}\rangle$. This being so, the expectation value of the stress tensor (14.90) is that of an isotropic fluid, with the pressure and density given by

$$\rho = \left\langle \tfrac{1}{2}\dot{\phi}^2 + \tfrac{1}{2}a^{-2}\nabla\phi\cdot\nabla\phi + V(\phi)\right\rangle \tag{14.92}$$

$$p = \left\langle \tfrac{1}{2}\dot{\phi}^2 - \tfrac{1}{6}a^{-2}\nabla\phi\cdot\nabla\phi - V(\phi)\right\rangle. \tag{14.93}$$

From these equations, we can see how a period of exponential expansion might come about. If the potential $V(\phi)$ were zero, we would have a field theory of free massless particles, whose energy density and pressure are given by the expectation values of the derivative terms in (14.92) and (14.93). In thermal equilibrium, at least, these contributions are proportional to T^4, and they fall as the universe expands. (It is often said that the kinetic energy of these particles is 'redshifted away'.) If we reach a state in which the energy density and pressure are mainly determined by the potential energy $V(\phi)$, then we have approximately $\rho = -p = \langle V(\phi)\rangle$. The stress tensor has approximately the form $T_{\mu\nu} = \langle V(\phi)\rangle g_{\mu\nu}$ and we see from the field equations (4.17) or from (14.18) and (14.19) that this 'vacuum energy' is in effect equivalent to a cosmological constant $\Lambda_{\text{eff}} = \kappa\langle V(\phi)\rangle$. The Friedmann equation becomes

$$\dot{a}^2 + k = \tfrac{1}{3}\Lambda_{\text{eff}}a^2. \tag{14.94}$$

Suppose that this equation first becomes approximately true at a time t_{i} when the scale factor is a_{i} and that for some period of time thereafter Λ_{eff} is approximately independent of time. For a flat universe, with $k = 0$, the solution is

$$a(t) = a_{\text{i}} \exp\left[(\tfrac{1}{3}\Lambda_{\text{eff}})^{1/2}(t - t_{\text{i}})\right]. \tag{14.95}$$

The cosmological model in which this is always true (that is, in which there is a cosmological constant but no ordinary matter) is called the *de Sitter* model. Because it has no matter, it is not a good model for our universe. The exponential expansion is much faster than the $t^{1/2}$ expansion envisaged in the radiation-dominated phase of the standard model. If such a period of *inflation* lasts long enough, then $a(t)$ can increase by a very large factor.

If the potential is the one with which our discussion began, namely $V(\phi) = (\lambda/4!)(\phi^2 - v^2)^2$, and the state preceding inflation is the one we envisaged with $\bar{\phi} = 0$, then we might guess that the effective cosmological constant during inflation is roughly $\Lambda_{\text{eff}} \approx \kappa V(0) = \kappa\lambda v^4/4!$. This low-temperature state, with a small density of particles, in which $\bar{\phi}$ is far from the minima at $\phi = \pm v$, is sometimes called a *false vacuum* state. Inflation persists while the energy density is dominated by the vacuum energy which means, in the example at hand, while $\bar{\phi}$ is close to zero. This false vacuum state is, however, unstable. We would expect it to evolve into a broken-symmetry state with, say, $\bar{\phi} = v$. A feature of this process that runs counter to normal intuition is that the effective cosmological constant corresponds roughly to a constant *energy density*, so that the total energy of the universe increases in proportion to a^3. As the broken-symmetry state emerges, this potential energy of the false vacuum must be converted into particles and radiation at a temperature comparable with, but somewhat lower than the critical temperature T_{c}—a process called *reheating*. From that point onwards, the history of the universe would be that described by the standard model.

The behaviour of the scale factor in the inflationary and standard models is sketched in figure 14.5, where I have simplified matters by supposing that inflation occurred more or less instantaneously at a time t_{I}, and that conditions were exactly the same just before inflation as they were just after, except that

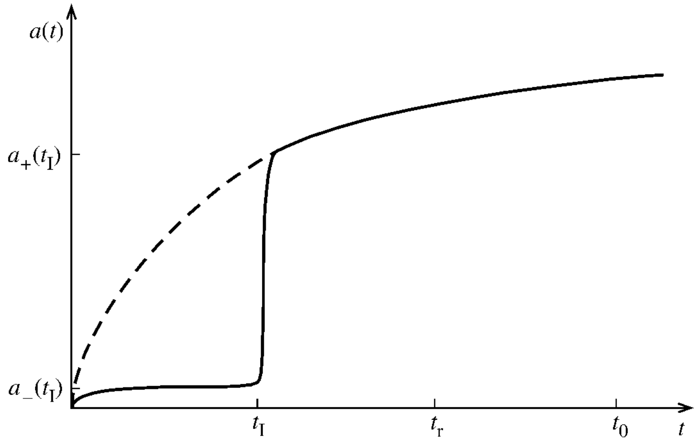


Figure 14.5. Schematic comparison of scale factors in the standard model (broken curve) and some versions of the inflationary model (solid curve). Neither the amount of inflationary expansion nor the relative time intervals is drawn to scale.

the size of a given comoving region was smaller by a factor $Z = a_+(t_\mathrm{I})/a_-(t_\mathrm{I})$. This implies that both models extrapolate backwards to an initial singularity at the same instant $t = 0$. It should be clear that inflation can solve the horizon problem if Z is sufficiently large. During the period before inflation, two small regions from which we now receive background radiation were much closer together than is allowed for in the standard model and could, after all, have communicated with each other. Let us see how large the factor Z has to be. The coordinate size of a region that could have become causally connected by the time t_I is given by an obvious modification of (14.72), namely

$$d = 2t_\mathrm{I}/a_-(t_\mathrm{I}) \tag{14.96}$$

and the coordinate distance D that a photon has travelled towards us since recombination is still given by (14.74). To solve the horizon problem, we need $d \geq D$, so that the entire observable universe lies within one causally connected region. (The extra distance that a causal influence could have travelled between t_I and t_r is essentially the same as the d that now subtends an angle of 1° and is too small to matter.) We can estimate the ratio d/D as

$$\begin{aligned}\frac{d}{D} &= \frac{2t_\mathrm{I}}{a_-(t_\mathrm{I})}\,\frac{a(t_\mathrm{r})}{3t_\mathrm{r}}\left[\frac{a(t_\mathrm{r})}{a(t_0)}\right]^{1/2}\\ &= \frac{2}{3}Z\frac{t_\mathrm{I}}{t_\mathrm{r}}\,\frac{a(t_\mathrm{r})}{a_+(t_\mathrm{I})}\left[\frac{a(t_\mathrm{r})}{a(t_0)}\right]^{1/2}\\ &= \frac{2}{3}Z\left(\frac{T_\mathrm{r}}{T_\mathrm{I}}\right)\left(\frac{T_0}{T_\mathrm{r}}\right)^{1/2}\end{aligned} \tag{14.97}$$

where I have assumed that the post-inflationary universe is radiation dominated, so that $T \propto a^{-1} \propto t^{-1/2}$ until $t = t_{\mathrm{r}}$. If we take $T_{\mathrm{I}} \sim 10^{15}\,\mathrm{GeV}/k_{\mathrm{B}} \sim 10^{28}$ K, corresponding roughly to the energy scale of grand unification, then we find $Z \gtrsim 10^{26} \sim \mathrm{e}^{60}$, which is usually expressed by saying that about 60 'e-folds' of the scale factor are needed. Evidently, figure 14.5 is not quite drawn to scale! A rough idea of how long it might take for the scale factor to increase by this amount can be gained by taking v to be, in energy units, about 10^{15} GeV and λ to be about 1. In laboratory units, the quantity $\frac{1}{3}\Lambda_{\mathrm{eff}}$ in (14.95) must be measured in s^{-2} so, inserting the appropriate factors of c and $\hbar$, we find that the required time interval is

$$\Delta t \approx 60\left(\tfrac{1}{3}\Lambda_{\mathrm{eff}}\right)^{-1/2} \approx 60\left(\frac{8\pi G v^4}{3\hbar^3 c^5}\right)^{-1/2} \approx 10^{-34}\,\mathrm{s}. \tag{14.98}$$

While these values of Z and Δt are fairly representative of the sort of numbers one encounters, the actual values depend somewhat on details of the theoretical models that are used and the assumptions that are introduced to deal with them.

To see how inflation can solve the flatness problem, we must solve the Friedmann equation (14.94) with $k = \pm 1$. The solution, with an initial scale factor $a_{\mathrm{i}} = a_{-}(t_{\mathrm{I}})$, is

$$\begin{aligned} a(t) = a_{\mathrm{i}} \cosh&\left[(\tfrac{1}{3}\Lambda_{\mathrm{eff}})^{1/2}(t - t_{\mathrm{i}})\right] \\ &+ \left(a_{\mathrm{i}}^2 - 3k/\Lambda_{\mathrm{eff}}\right)^{1/2} \sinh\left[(\tfrac{1}{3}\Lambda_{\mathrm{eff}})^{1/2}(t - t_{\mathrm{i}})\right]. \end{aligned} \tag{14.99}$$

For large values of their argument, both $\cosh\theta$ and $\sinh\theta$ are approximately equal to $\frac{1}{2}\exp\theta$, so if (14.94) is valid for a period of time longer than about $(\frac{1}{3}\Lambda_{\mathrm{eff}})^{-1/2}$ we again have exponential expansion. During this expansion, the Hubble parameter $H = \dot{a}/a$ is just a constant, equal to $(\frac{1}{3}\Lambda_{\mathrm{eff}})^{1/2}$. If a itself becomes very large, then k/a^2 becomes negligible compared with H^2 and the universe is very close to being flat. Intuitively, we may imagine, for example, a balloon inflated to a very large size. The part of the universe we observe corresponds to a tiny fraction of its surface, which will appear almost flat. At the end of inflation, the potential energy density $\Lambda_{\mathrm{eff}}/\kappa$ is converted into an equivalent energy density in particles and radiation, which is automatically equal to the critical density $3H^2/\kappa$. If the part of the universe that we can observe has once been made flat to a high degree of accuracy by this mechanism, then it remains flat. That is to say, the term k/a^2 in (14.18) and (14.19) remains negligible, and the function $a(t)$ that solves these equations automatically leads to (14.38), regardless of how the effective density ratio Ω may be made up from baryonic matter, radiation, non-baryonic matter and a cosmological constant.

The question naturally arises, whether the sequence of events that I have outlined really does result from the solution of the field equations (14.91) when the stress tensor is that of a quantum field theory that might reasonably be thought to describe the matter in our universe. The problem of calculating $\langle T_{\mu\nu}\rangle$ for even a simple quantum field theory in a non-equilibrium state proves to be extremely difficult (see, for example, Lawrie (1999)) and such calculations have been attempted only for models that are too highly idealized for any firm conclusions to be drawn. The strategy most often adopted by cosmologists is to assume that the non-equilibrium state of the quantum field can adequately be characterized by the value of a *classical* scalar field, which has a definite value at each point of spacetime. In a homogeneous universe, this value can depend only on the cosmic time t, so the energy density and pressure are just

$$\rho = \tfrac{1}{2}\dot{\phi}^2 + V(\phi) \qquad p = \tfrac{1}{2}\dot{\phi}^2 - V(\phi). \tag{14.100}$$

If these expressions are substituted into (14.21) and (14.22), a short calculation shows that the equation of motion for ϕ itself must be

$$\ddot{\phi} + 3H\dot{\phi} = -V'(\phi) \tag{14.101}$$

where $V'(\phi) = \mathrm{d}V(\phi)/\mathrm{d}\phi$ and $H = \dot{a}/a$. In fact, the Euler–Lagrange equation obtained from the action (14.89) is (exercise 14.3)

$$\ddot{\phi} + 3H\dot{\phi} - \frac{1}{a^2}\nabla^2\phi + V'(\phi) = 0 \tag{14.102}$$

and this, of course, reduces to (14.101) when ϕ depends only on t. This equation has the same form as the equation for a Newtonian particle whose position in a one-dimensional space is ϕ and whose potential energy is $V(\phi)$, if we imagine this particle also to be subject to a frictional force $-3H\dot{\phi}$.

With the reasonable assumption that it is sufficient to deal with a region of the universe that can be considered flat, the Friedmann equation (14.21) now becomes

$$H^2 = \tfrac{1}{3}\kappa\left[\tfrac{1}{2}\dot{\phi}^2 + V(\phi)\right]. \tag{14.103}$$

Within this scheme, the equation of motion (14.101) for ϕ and the Friedmann equation (14.103) form a closed set, which can be solved (numerically, if not analytically) to find the evolution of the universe from a given initial state. The question whether inflation can occur can be addressed in a preliminary way without a detailed solution, however. A minimal requirement is that the expansion should accelerate, which means that

$$\ddot{a}/a = -\tfrac{1}{6}\kappa(\rho + 3p) = -\tfrac{1}{3}\kappa[\dot{\phi}^2 - V(\phi)] > 0. \tag{14.104}$$

Consider the slightly stronger requirement that

$$\dot{\phi}^2 \ll V(\phi) \qquad \text{which implies} \qquad H^2 \approx \tfrac{1}{3}\kappa V(\phi). \tag{14.105}$$

Supposing that this condition is to hold over some significant period of time, then it should also be true that $\mathrm{d}\dot{\phi}^2/\mathrm{d}t \ll \mathrm{d}V(\phi)/\mathrm{d}t$, or

$$\ddot{\phi} \ll V'(\phi). \tag{14.106}$$

If so, then the term $\ddot{\phi}$ can be neglected in (14.101), with the result that

$$3H\dot{\phi} \approx -V'(\phi). \tag{14.107}$$

The analogue Newtonian particle, that is to say, has reached a 'terminal velocity', such that the frictional force balances the potential gradient. If the expansion is to be approximately exponential, then H must be approximately constant, so we can differentiate (14.107) to find

$$\ddot{\phi} \approx -(3H)^{-1}V''(\phi)\dot{\phi}. \tag{14.108}$$

With a little rearrangement, the two conditions (14.105) and (14.106) become

$$\left|\frac{V''(\phi)}{V(\phi)}\right| \ll 3\kappa \qquad \text{and} \qquad \left|\frac{V'(\phi)}{V(\phi)}\right| \ll \sqrt{3\kappa}. \tag{14.109}$$

These are restrictions on the shape of the potential $V(\phi)$, which tell us that it must, for some range of values of ϕ, be rather flat. They are sufficient (though not strictly necessary) conditions for the occurrence of some period of inflation; when they are met, the jargon has it that a 'slow roll' approximation applies. Suppose that ϕ traverses a range of values from ϕ_1 to ϕ_2 where these conditions are satisfied. The scale factor can be written as $a = a_0 \exp\left(\int H(t)\mathrm{d}t\right)$, where a_0 is a constant, so we can use (14.105) and (14.107) to estimate the number of e-folds as

$$n_\mathrm{e} = \int_{t_1}^{t_2} H(t)\mathrm{d}t = \int_{\phi_1}^{\phi_2} \frac{H}{\dot{\phi}}\mathrm{d}\phi = -\kappa \int_{\phi_1}^{\phi_2} \frac{V(\phi)}{V'(\phi)}\,\mathrm{d}\phi. \tag{14.110}$$

Essentially this 'slow roll' idea, which differs in some important details from Guth's original proposal, was first deployed by Linde (1982) and by Albrecht and Steinhardt (1982) in connection with the phase transition in a grand-unified theory, from which our discussion started. The potential $V(\phi)$ they considered was not the one that appears in the Lagrangian of the theory, but rather an *effective potential*, calculated roughly in the same way as (14.86), but including additional corrections that arise from the interaction of ϕ with the gauge bosons. If the parameters in the theory are appropriately chosen, this potential has roughly the form sketched in figure 14.6. There are regions in which it is very flat, as required by (14.109), and it is found that an expansion factor Z much greater than the required value of 10^{26} or so is possible. Nevertheless, cosmologists are generally agreed that this mechanism does not work. The reason lies in what has come to be the most prominent feature of inflationary cosmology, namely a prediction of small perturbations, or inhomogeneities, in the energy density ρ. The presence of inhomogeneities in the fairly early universe is necessary to account for the presently observed clumping of matter into galaxies and clusters of galaxies (the umbrella term for which is *large-scale structure*). Structure of the general sort that we now see can be shown to come about through the gravitational

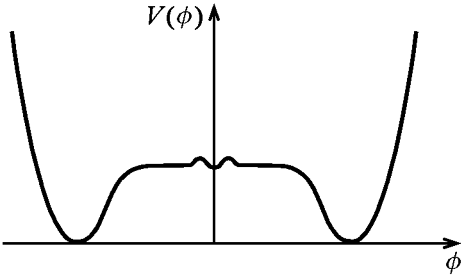


Figure 14.6. Qualitative form of the effective potential assumed in some versions of the inflationary model.

attraction of regions whose density may initially have been only slightly greater than the average, and cosmologists have devised methods of studying this process in great detail. Small inhomogeneities in the density of matter at the time of recombination would be reflected in variations, on small angular scales, of the temperature of the microwave background, and these have indeed been observed at the level of about one part in 10^5.

The prevailing view among inflation theorists is that these inhomogeneities have a quantum-mechanical origin. Any inhomogeneities that existed prior to inflation would have been smoothed out by the inflationary expansion, so those that are relevant to observations were created while inflation was taking place. During this period, the energy density was entirely dominated by the potential energy $V(\phi)$, so a mechanism that created small, inhomogeneous fluctuations in $\phi(t)$, say $\varphi(\boldsymbol{x},t)$, would lead to corresponding perturbations in the density $\delta\rho(\boldsymbol{x},t) = V'(\phi)\varphi(\boldsymbol{x},t)$. This expression does not give us directly the perturbations that would have been present at, say, the time of recombination, because conditions in the universe would have evolved significantly between these two times. It proves possible, however, to estimate the density perturbations in the radiation-dominated 'fairly early' universe without the need to know exactly what happened in the intervening period. The following argument (a simplified version of one due to Guth and Pi (1982) and to Starobinski (1982)) indicates how this can be done, but side-steps several questions which must be dealt with in a more complete analysis.

We express the inhomogeneous scalar field as

$$\phi(\boldsymbol{x},t) = \phi_0(t) + \varphi(\boldsymbol{x},t) \tag{14.111}$$

where the average field $\phi_0(t)$ obeys (14.101). The whole field $\phi(\boldsymbol{x},t)$ obeys (14.102), but we linearize this equation, assuming that $\varphi(\boldsymbol{x},t)$ is small. The result is

$$\ddot{\varphi} + 3H\dot{\varphi} - a^{-2}\nabla^2\varphi + V''(\phi_0)\varphi = 0 \tag{14.112}$$

and it will be helpful to compare this with the equation satisfied by $\dot{\phi}_0(t)$, obtained by differentiating (14.101). With the assumption that H is approximately constant during inflation, we find

$$\partial_t^2\dot{\phi}_0 + 3H\,\partial_t\dot{\phi}_0 + V''(\phi_0)\dot{\phi}_0 = 0. \tag{14.113}$$

Suppose that the term $a^{-2}\nabla^2\varphi$ can be neglected. Then these two equations are identical and the time dependence of $\varphi(\boldsymbol{x}, t)$ is the same as that of $\dot{\phi}_0(t)$. Later on, we shall have to determine just when this is true, but for now I simply assume that it is. Then we can write

$$\varphi(\boldsymbol{x}, t) = -\tau(\boldsymbol{x})\dot{\phi}_0(t) \tag{14.114}$$

where $\tau(\boldsymbol{x})$ is a small, time-independent function, and up to corrections of order φ^2 we have

$$\phi(\boldsymbol{x}, t) \approx \phi_0(t) - \tau(\boldsymbol{x})\dot{\phi}_0(t) \approx \phi_0\Big(t - \tau(\boldsymbol{x})\Big). \tag{14.115}$$

Thus, the net effect of the perturbation is a position-dependent time delay (which might be either positive or negative) in the 'rolling' of ϕ.

At this point, we must recognize that the splitting of spacetime into spatial sections corresponding to definite values of t, which was natural in terms of the exact Robertson–Walker metric, is now slightly ambiguous. Let us, indeed, define a new time coordinate $\bar{t} = t - \tau(\boldsymbol{x})$. On the constant-$\bar{t}$ sections of spacetime, the field $\phi(\bar{t})$ and the density $\rho(\bar{t})$ are constant, but the scale factor varies:

$$\bar{a}(\boldsymbol{x}, \bar{t}) = a\Big(\bar{t} + \tau(\boldsymbol{x})\Big) \approx a(\bar{t}) + \tau(\boldsymbol{x})\dot{a}(\bar{t}) \approx [1 + H_{\mathrm{I}}\tau(\boldsymbol{x})]\,a(\bar{t}) \tag{14.116}$$

where H_{I} is the roughly constant value of the Hubble parameter during inflation. When inflation comes to an end, the vacuum energy density is converted into particles and radiation. Initially, this density is uniform over the spatial sections of constant $\bar{t}$, but in the radiation-dominated era it varies as $1/a^4$. Assuming that each small region evolves like a miniature Friedmann–Robertson–Walker universe, with its own scale factor, the density at later times will be

$$\rho(\boldsymbol{x}, \bar{t}) = \frac{\rho(\bar{t})}{[1 + H_{\mathrm{I}}\tau(\boldsymbol{x})]^4} \simeq \rho(\bar{t})\left[1 + \delta_\rho(\boldsymbol{x})\right] \tag{14.117}$$

where $\rho(\bar{t}) \simeq \text{constant}/a^4(\bar{t})$ and the fractional variation in density, or *density contrast*, is given by

$$\delta_\rho(\boldsymbol{x}) = \frac{\delta\rho(\boldsymbol{x})}{\rho} = -4H_{\mathrm{I}}\tau(\boldsymbol{x}). \tag{14.118}$$

Wary readers will appreciate that this simple calculation has many pitfalls. Chief among them is the fact that the variations in density over a 'spatial' section of spacetime depend on how that spatial section is chosen; we could apparently manufacture arbitrary density perturbations by choosing a new time coordinate

$t' = \bar{t} + \delta t(\boldsymbol{x})$ and redefining 'space' to be the three-dimensional surface of constant t'. The dependence of $\rho(\boldsymbol{x}, t)$ on our choice of a coordinate system is usually referred to as a 'gauge' dependence, by analogy with the gauge degrees of freedom in electromagnetism and other gauge theories, and only gauge-independent quantities have a genuine physical meaning. By means of a sufficiently careful analysis, it is possible to arrive at predictions for observed variations in, for example, the microwave background radiation that are free of gauge ambiguities, and the result is substantially equivalent to (14.118). This gauge-invariant analysis can be developed in several different ways, all of which are too lengthy for me to enter into them here. Readers who would like to know more about them might start with the account given by Liddle and Lyth (2000).

It remains to estimate the size of $\tau(\boldsymbol{x})$ and to this end it is useful to take Fourier transforms

$$\varphi(\boldsymbol{x}, t) = \int \frac{\mathrm{d}^3 k}{(2\pi)^3}\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \varphi_{\boldsymbol{k}}(t) \tag{14.119}$$

and similarly for $\tau(\boldsymbol{x})$ and $\delta_\rho(\boldsymbol{x})$. So long as we deal only with perturbations in both the density and the metric that are linear in $\varphi(\boldsymbol{x}, t)$, each Fourier component evolves independently of the others. From (14.112) we obtain

$$\ddot{\varphi}_{\boldsymbol{k}} + 3H\dot{\varphi}_{\boldsymbol{k}} + (k/a)^2\varphi_{\boldsymbol{k}} + V''(\phi_0)\varphi_{\boldsymbol{k}} = 0 \tag{14.120}$$

where $k = |\boldsymbol{k}|$. The calculation that led to (14.118) assumed that the third term, which has become $(k/a)^2\varphi_{\boldsymbol{k}}$, was negligible, and it will now be important to find out when this is true. Let us write $\varphi_{\boldsymbol{k}}(t) = \exp[f_{\boldsymbol{k}}(t)]$, so that the equation of motion (14.120) becomes

$$\ddot{f}_{\boldsymbol{k}} + \dot{f}_{\boldsymbol{k}}^2 + 3H\dot{f}_{\boldsymbol{k}} + (k/a)^2 + V''(\phi_0) = 0. \tag{14.121}$$

Consider the trial solution $\dot{f}_{\boldsymbol{k}} \approx -3H$. As before, we take H to be roughly constant, which implies that $\ddot{f}_{\boldsymbol{k}} \approx 0$. This solution will be approximately valid if the last two terms in (14.121) are much smaller than $\dot{f}_{\boldsymbol{k}}^2$ and $|3H\dot{f}_{\boldsymbol{k}}|$, both of which are equal to $9H^2$. The criteria for this are

$$(k/Ha)^2 \ll 9 \tag{14.122}$$

$$|V''(\phi_0)/H^2| \ll 9. \tag{14.123}$$

Now the 'slow roll' conditions (14.105) and (14.109) imply $|V''/H^2| \ll 9$, which is precisely (14.123), so the substantive criterion is (14.122). In this inequality, the wave vector k is (up to a factor of 2π) the inverse wavelength of the perturbation as measured by the comoving coordinates $\boldsymbol{x}$, so a/k is the physical wavelength, often referred to as the length scale, or simply the 'scale' of the perturbation. In this context, H^{-1} is a characteristic distance in the inflating (or de Sitter) spacetime. It is often called the 'horizon' because, according to Hubble's law (14.9), it is the separation of two points that are being carried apart at the speed of light, and is therefore the greatest distance over which any causal influence might act. The jargon has it that a scale for which $a/k < H^{-1}$ is 'inside the horizon', while one for which $a/k > H$ is outside. Roughly speaking, then, for scales which are outside the horizon, the equation of motion (14.120) can be replaced by $\ddot{\varphi}_{\boldsymbol{k}} + 3H\dot{\varphi}_{\boldsymbol{k}} \approx 0$, whose solution is

$$\varphi_{\boldsymbol{k}}(t) \approx \varphi_{\boldsymbol{k}} + b_{\boldsymbol{k}}\mathrm{e}^{-3Ht} \tag{14.124}$$

where $\varphi_{\boldsymbol{k}}$ and $b_{\boldsymbol{k}}$ are constants. Since a/k increases exponentially with time, the net result is that each Fourier component eventually 'crosses the horizon' and settles to a constant value $\varphi_{\boldsymbol{k}}$ shortly thereafter, say at $t \simeq t^*(k)$. According to (14.118) and (14.114), we therefore estimate

$$|\delta_\rho(\boldsymbol{k})| \approx 4H_{\mathrm{I}}|\tau_{\boldsymbol{k}}| \approx 4H_{\mathrm{I}}\left|\frac{\varphi_{\boldsymbol{k}}}{\dot{\phi}_0(t^*)}\right|. \tag{14.125}$$

Finally, we need an estimate of $\varphi_{\boldsymbol{k}}$. The conventional strategy is to reinstate the quantum-mechanical nature of the field $\phi(\boldsymbol{x}, t)$, taking $\varphi(\boldsymbol{x}, t)$ to be a free quantum field with the equation of motion

$$\ddot{\varphi} + 3H\dot{\varphi} - a^{-2}\nabla^2\varphi = 0. \tag{14.126}$$

In (14.125), the quantity $\varphi_{\boldsymbol{k}}$ is to be identified as a root-mean-square average over the quantum indeterminacy in $\varphi(\boldsymbol{x}, t)$. To be specific, we define the *power spectrum* for density perturbations by

$$P_\rho(k) = 4\pi k^3\langle\delta_\rho^2(\boldsymbol{k})\rangle \tag{14.127}$$

and identify

$$\langle\varphi_{\boldsymbol{k}}^2\rangle = (2\pi)^{-3}\int \mathrm{d}^3x\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\langle 0|\varphi(\boldsymbol{x}, t)\varphi(\boldsymbol{0}, t)|0\rangle \tag{14.128}$$

where $|0\rangle$ is a vacuum state for φ. This expectation value can be calculated as follows. The action for φ is (14.89) with $V = 0$. Writing the metric explicitly in terms of the scale factor, we have

$$S = \int \mathrm{d}^4x\, a^3\left[\tfrac{1}{2}\dot{\varphi}^2 - \tfrac{1}{2}a^{-2}\nabla\varphi\cdot\nabla\varphi\right] \tag{14.129}$$

which yields the momentum conjugate to φ as $\Pi(\boldsymbol{x}, t) = a^3(t)\dot{\varphi}(\boldsymbol{x}, t)$. The general solution of (14.126) analogous to the solution (7.11) of the Klein–Gordon equation in Minkowski spacetime can be written as

$$\varphi(\boldsymbol{x}, t) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\left[\alpha_{\boldsymbol{k}}\chi_{\boldsymbol{k}}(t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} + \alpha_{\boldsymbol{k}}^\dagger\chi_{\boldsymbol{k}}^*(t)\mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\right] \tag{14.130}$$

where $\alpha_{\boldsymbol{k}}$ and $\alpha_{\boldsymbol{k}}^\dagger$ are creation and annihilation operators with the commutator

$$\left[\alpha_{\boldsymbol{k}}, \alpha_{\boldsymbol{k}'}^\dagger\right] = (2\pi)^3\delta(\boldsymbol{k} - \boldsymbol{k}') \tag{14.131}$$

and $\chi_{\boldsymbol{k}}(t)$ is a solution of (14.120) with $V'' = 0$. The important point is that $\chi_{\boldsymbol{k}}(t)$ should have the correct magnitude. This is determined by the commutation relation

$$\left[\varphi(\boldsymbol{x},t),\dot{\varphi}(\boldsymbol{x}',t)\right] = a^{-3}(t)\left[\varphi(\boldsymbol{x},t),\Pi(\boldsymbol{x}',t)\right] = \mathrm{i}a^{-3}(t)\delta(\boldsymbol{x}-\boldsymbol{x}') \quad (14.132)$$

which will be true if $\chi(\boldsymbol{x},t)$ satisfies the *Wronskian condition*

$$\chi_{\boldsymbol{k}}^{*}(t)\dot{\chi}_{\boldsymbol{k}}(t) - \dot{\chi}_{\boldsymbol{k}}^{*}(t)\chi_{\boldsymbol{k}}(t) = -\mathrm{i}a^{-3}(t). \quad (14.133)$$

A suitable function is

$$\chi_{\boldsymbol{k}}(t) = \frac{H}{\left(2k^3\right)^{1/2}}\left(1-\frac{\mathrm{i}k}{aH}\right)\exp\left(\frac{\mathrm{i}k}{aH}\right). \quad (14.134)$$

The state $|0\rangle$ is defined by $\alpha_{\boldsymbol{k}}|0\rangle = \langle 0|\alpha_{\boldsymbol{k}}^{\dagger} = 0$ and it is straightforward to calculate the expectation value, which is

$$\langle\varphi_{\boldsymbol{k}}^2\rangle = \frac{H^2}{2(2\pi)^3k^3}\left(1+\frac{k^2}{a^2H^2}\right) \simeq \frac{H^2}{2(2\pi)^3k^3} \quad (14.135)$$

In the second expression, I have taken $k/aH \ll 1$ for a perturbation outside the horizon.

Altogether, the power spectrum turns out to be

$$P_\rho(k) \sim \left(\frac{H_\mathrm{I}^2}{2\pi\dot{\phi}_0(t^*)}\right)^2 \quad (14.136)$$

where $\sim$ indicates that there is a numerical factor of order 1, which depends somewhat on the details of how the estimate is made. Again, H_I is the Hubble parameter of the inflationary universe although the density perturbations described by this power spectrum are those in the radiation-dominated era. It should be mentioned, though, that this power spectrum is not directly observed. In the radiation-dominated era, the scale factor varies as $a \propto t^{1/2}$ and the Hubble parameter as $H \propto t^{-1}$, so the ratio $k/(aH)$ increases as $t^{1/2}$. Thus, length scales which moved outside the horizon during inflation may re-enter the horizon during the radiation-dominated era. When they do, causal processes may again affect the evolution of perturbations on these length scales. A well-established theory indicates that $\delta_\rho(t)$ then oscillates with time. The temperature variations in the microwave background, when determined with enough precision as a function of length scale, should reveal what amounts to a snapshot of these oscillations at the time of recombination (or on the 'surface of last scattering', as the matter is usually put). Given the primordial power spectrum (14.136), the expected temperature variations can be determined reliably, as long as the perturbations are not so large as to invalidate the use of linearized equations of motion.

A notable feature of the power spectrum is that it is substantially independent of k, although $\dot{\phi}_0\left(t^*(k)\right)$ does depend weakly on k. It is said to be approximately *scale invariant*. Long before the idea of inflation was conceived, it had been argued by E R Harrison and Ya B Zeldovich that a scale-invariant spectrum of density perturbations was needed to account for the subsequent development of galactic clusters, and that the magnitude of δ_ρ should be of the order of 10^{-5}–10^{-4}, which is consistent with the temperature fluctuations in the microwave background observed subsequently. This, of course, is a significant point in favour of the inflationary account.

Unfortunately, the magnitude of δ_ρ as calculated from the potential $V(\phi)$ of a grand unified theory turns out to be too large by a factor of around 10^5. For this reason, most cosmologists no longer regard as tenable the idea of inflation arising from a phase transition associated with grand unification. More recent developments have largely been based on the idea of *chaotic inflation* (Linde, 1983). When the universe was young enough for typical energies to have been of the order of the Planck energy (about 10^{19} GeV—see appendix C) or above, the poorly understood effects of quantum gravity are likely to have been important. According to Linde, one might expect the universe to have emerged from the quantum gravity era in a chaotic state. In particular, the value of the scalar field ϕ would vary widely from one region of the universe to another and in some region, destined to become the one that we now observe, would have had the value needed for one's favourite inflationary scenario to work.

The popular pastime of inflationary model-building has produced an enormous variety of models, whose virtues and shortcomings I cannot usefully survey here; interested readers may like to consult Liddle and Lyth (2000). Just what can be learned from these models is, I confess, something that I find it hard to assess. Typically, to reduce the magnitude of density perturbations to an acceptable level, it is necessary for ϕ to be very weakly coupled to other fields, so it cannot be identified as a Higgs field and is generally referred to as the *inflaton*. Most often, indeed, the inflaton and its (arbitrarily adjustable) potential are invoked solely for the purpose of producing inflation, and have nothing to do with any established particle physics, although speculative theories of supersymmetry, supergravity and superstrings provide motivations for some of the variants that have been proposed. An optimistic point of view is that sufficiently detailed agreement between predictions for the density perturbations and (future) precise measurements might allow one to infer the existence of the requisite inflaton and the form of its potential. However, in the absence of other corroborating evidence, such as the identification of a particle species that might plausibly be associated with the inflaton field, it would seem difficult to argue that the observed density perturbations might not have some quite different origin.

It is also possible to wonder whether calculations of the kind that I have sketched above really have a secure basis in quantum field theory. For example, the classical field $\phi(\boldsymbol{x}, t)$ is often represented as being the expectation value of a quantum field, but this cannot be exactly right. According to the standard interpretation of quantum mechanics, the expectation value is the expected average of repeated measurements made on many identically prepared systems, but such repeated measurements are not easy to carry out on an ensemble of identically prepared universes. In any case, the things that one might, in principle, regard as measurable are quantities such as energy density and pressure, and a definite result from such a measurement does not entail a definite value for an underlying object such as ϕ. Moreover, the expectation value of the stress tensor, $\langle T_{\mu\nu}(\phi)\rangle$ that appears in (14.91) is by no means the same thing as $T_{\mu\nu}(\langle\phi\rangle)$ and in general it cannot be expressed as a function of the single variable $\langle\phi\rangle$ at all; many more parameters are needed to characterize the state of a non-equilibrium quantum system. I pointed out that the semiclassical field equations (14.91) themselves cannot be correct at a fundamental level, so one might wonder whether replacing all quantum fields by their expectation values might be an equally plausible strategy. That it would not is clear from the fact that this would yield $\rho = p = 0$ in the case of a radiation-dominated universe containing only photons. In the same way, one may wonder about the legitimacy of treating the quantum *indeterminacy* of a field such as ϕ as being equivalent to a real fluctuation in a classical energy density. Questions such as these have been asked from time to time (see, for example, Evans and McCarthy (1985), Guth and Pi (1985), Mazenko *et al* (1985), Albrecht *et al* (1994), Boyanovsky *et al* (1998)). For what it is worth, my own (possibly eccentric) view is that they have not been convincingly answered, but readers who pursue for themselves the discussions to be found in the literature may well arrive at a different conclusion.

Exercises

14.1. The *absolute luminosity* L of an astronomical object is the total power it radiates. Its *apparent luminosity* ℓ is the power per unit area received by an observer. In Euclidean space, the apparent luminosity for an observer at a distance d is obviously $\ell = L/4\pi d^2$. In general, the *luminosity distance* of a source of known luminosity is defined as $d_{\mathrm{L}} = (L/4\pi\ell)^{1/2}$. Consider a comoving source and a comoving observer separated by a coordinate distance r in a Robertson–Walker spacetime. Radiation emitted at time t_{e} is received at time t_0. By considering both the rate at which photons are received and the redshift of each photon, show that

$$\ell = \frac{La^2(t_{\mathrm{e}})}{4\pi r^2 a^4(t_0)}.$$

The scale factor at time t can be expressed as a power series in $(t - t_0)$ as

$$a(t) = a(t_0)\left[1 + H_0(t - t_0) - \tfrac{1}{2}q_0 H_0^2(t - t_0)^2 + \ldots\right].$$

Use this expansion and (14.14) to express the redshift z and the coordinate distance r as power series in $(t_0 - t_{\mathrm{e}})$ and hence express r as a power series in z. Show that the luminosity distance is given by (14.11).

14.2. The covariant action for a massless, conformally coupled scalar field can be written as

$$S = \tfrac{1}{2}\int \mathrm{d}^4x\,(-g)^{1/2}[g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + \xi R\phi^2]$$

with $\xi = \frac{1}{6}$. Considering a spatially flat Robertson–Walker spacetime, for which the Ricci scalar R is given by (14.7), and using Cartesian spatial coordinates, derive the Euler–Lagrange equation. Show that it has plane-wave solutions of the form

$$\phi(\boldsymbol{x}, t) = [2\omega(t)a^3(t)]^{-1/2}\exp\left(-\mathrm{i}\int_{t_0}^{t}\omega(t')\mathrm{d}t' + \mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}\right)$$

where the time dependent frequency satisfies the equation

$$\omega^2 + \frac{1}{2}\frac{\ddot{\omega}}{\omega} - \frac{3}{4}\frac{\dot{\omega}^2}{\omega^2} = \frac{|\boldsymbol{k}|^2}{a^2} - \frac{1}{2}\frac{\ddot{a}}{a} - \frac{1}{4}\frac{\dot{a}^2}{a^2}.$$

Verify that this equation is satisfied by $\omega(t) = |\boldsymbol{k}|/a(t)$ and hence that the frequency and wavelength of the particle are redshifted as in (14.16).

14.3. By adding to the action of the previous exercise a potential $V(\phi)$, and setting $\xi = 0$, deduce the equation of motion (14.102).

14.4. Consider a projectile launched vertically from the surface of the Earth. Write down an expression for its total energy E, with the usual convention that the potential energy vanishes at $r \to \infty$. The escape velocity corresponds to $E = 0$. Verify that the Friedmann equation (14.21) with $\rho = M/a^3$ has exactly the same form, with $k \propto -E$.

14.5. (a) For a radiation-dominated universe, show that the function $f(\Omega)$ in (14.34) is given by $f(\Omega) = 1/(1 + \Omega^{1/2})$.

(b) Consider a flat universe with a positive cosmological constant, containing only pressureless, non-relativistic matter. By integrating the Friedmann equation (14.18), show that

$$t = \frac{2}{(3\Lambda)^{1/2}}\ln\left[\frac{\Lambda^{1/2} + (\kappa\rho + \Lambda)^{1/2}}{(\kappa\rho)^{1/2}}\right]$$

and hence that the function $f(\Omega)$ is

$$f(\Omega) = \frac{2}{3(1-\Omega)^{1/2}}\ln\left[\frac{1 + (1-\Omega)^{1/2}}{\Omega^{1/2}}\right].$$

Verify that $f(1) = \frac{2}{3}$.

14.6. The discussion of §14.3 assumes a cosmological model in which $\rho(t) = \rho_{\mathrm{matter}}(t) + \rho_{\mathrm{rad}}(t) = M/a^3(t) + \Gamma/a^4(t)$, where M and Γ are constants, and

$p(t) = p_{\text{rad}}(t) = \Gamma/3a^4(t)$. Verify that such a model is consistent with both of equations (14.21) and (14.22).

14.7. With a positive cosmological constant Λ, show that a static universe (the *Einstein universe*) with a, ρ and p all constant is possible provided that $\rho \leq 2\Lambda/\kappa$, and that this universe is closed. In the *Lemaître* universe, p is taken to be zero and the constant $M = \rho a^3$ is larger than the value required for a static universe. Show that (i) this model has an initial singularity with $a(t)$ initially proportional to $t^{2/3}$; (ii) the expansion slows down until $\dot{a}$ reaches a minimum when $a^3 = \kappa M/2\Lambda$; (iii) after a sufficiently long time, the expansion becomes exponential as in the de Sitter universe (14.95).

Chapter 15

An Introduction to String Theory

At the end of chapter 12, we left the enterprise of constructing a unified theory of fundamental particles and their interactions in a rather unsatisfactory state. As judged by its ability to reproduce the observed phenomena of particle physics, the standard model is outstandingly successful, but it leaves many questions unanswered. There are twenty or so parameters (coupling constants and masses) whose values cannot be deduced from any principles of the theory and must simply be adjusted to fit the facts. Likewise, the gauge symmetry group SU(3)×SU(2)×U(1) and the number of families of quarks and leptons must be chosen, from the limitless possibilities that would seem to present themselves, just so as to fit the facts. The apparent convergence of the running coupling constants of the standard model to a single value at around 10^{15}–10^{16} GeV seems to point towards a more completely unified underlying theory. If this is taken to be a grand-unified gauge theory, though, then disturbingly *ad hoc* measures (such as the fine tuning of some of its parameters) are needed to fit the known facts, while other features of the theory (such as its gauge group) cannot be determined because not enough facts are known!

Moreover, none of these theories includes a description of gravitational forces. General relativity, although it cannot be tested quite as stringently as the standard model, is also a highly successful theory. As we discovered in chapter 8, it is somewhat akin to the gauge theories of particle physics, but all attempts to convert it into a quantum-mechanical theory have been unsuccessful. It is worth emphasizing that we cannot be satisfied with a classical theory of gravity. Quite apart from any aesthetic prejudice, the field equations (4.17) simply do not make sense if the geometrical tensor on the left-hand side is a classical one while the stress tensor on the right-hand side is a quantum-mechanical operator, as it must be. The combination of a classical theory of spacetime with a quantum theory of the matter that lives there does not lead to a self-consistent view of the world.

In the 1970s there emerged, more or less by accident, the beginnings of a theory which seems to offer the hope of a truly unified and self-consistent view of the world. Whether it is a correct view is quite another matter: there is currently no shred of experimental evidence that would serve either to confirm or to refute the mathematical notions that have been advanced. Its point of departure is the idea that the fundamental constituents of matter are not point particles but one-dimensional objects called, quite reasonably, *strings*. As I write these lines, some thirty years on, the theory has become so extensive that I cannot hope to do it justice in a single chapter. The greater part of this chapter is intended to give substance to three key ideas: (i) that the various particle species we observe might be identifiable as different states of vibration of a single basic object—the relativistic string; (ii) that one of these vibrational states can be identified as the graviton, and consequently that the theory does indeed include a quantum-mechanical description of spacetime geometry; (iii) that the quantum mechanics of a relativistic string requires (at least in the most usual version of the theory) the existence of more spacetime dimensions than the four that are familiar to us.

I shall develop in some detail the theory of a *free bosonic string*, whose only physical attributes are its location in (and motion through) spacetime. We begin in §15.1 by looking briefly at the quantum mechanics of a relativistic point particle from a point of view which is different from the one we have taken until now, but is more readily generalized to the case of a string. The classical theory of a relativistic string occupies §15.2, where we shall see that a tractable mathematical formalism involves physically redundant degrees of freedom analogous to the gauge degrees of freedom of electromagnetism and its non-Abelian generalizations. (In fact, they are very similar to the gauge degrees of freedom which, as we saw in §7.6.2, result from the coordinate invariance of general relativity.) The quantization of this classical theory is dealt with in §15.3. Although the basic procedure is the one familiar from chapter 5, we shall discover that very careful thought is needed to deal correctly with the gauge degrees of freedom. In fact, the quantization can be carried through consistently only on condition that the gauge symmetry of the classical theory survives as a symmetry of the quantum theory and it is this condition which, as we learn in §15.3.3, appears to demand the existence of extra spacetime dimensions. The physical interpretation of this prototype string theory is the subject of section §15.4, where we shall see how to construct the physical states that can be identified as particles of definite mass and spin. In particular, we find that one of these particles is a massless spin-2 particle, which we would like to identify as a graviton, and §15.4.3 shows how the existence of this state of the string is related to changes in spacetime geometry. Finally, I discuss much more qualitatively in §15.5 some of the further advances that have been made in the attempt to turn this prototype theory into a real working model of the physical world. These include the mechanism whereby strings may interact (§15.5.1); the supersymmetric strings (§15.5.2), whose additional degrees of freedom are needed to account for the existence of fermionic particles and of internal symmetries such the gauge symmetry of the standard model, as well as for more technical reasons of mathematical consistency; and some of the implications of the compactification of extra spacetime dimensions (§15.5.3). In the end, we shall see that the most recent developments point towards a still deeper theory, the exact nature of which is still unclear.

Unavoidably, this chapter will be somewhat more technical than most of its predecessors, and we shall have to work rather hard to obtain just a few key results. Even so, we shall be able only to scratch the surface of what has become a very large and mathematically sophisticated branch of theoretical physics.

15.1 The Relativistic Point Particle

We wrote down in (3.32) a Lagrangian for a free classical particle in Minkowski spacetime. In a slightly different notation, the corresponding action is

$$S = -\frac{m}{2}\int \mathrm{d}\tau_{\mathrm{p}}\,\frac{\mathrm{d}X_\mu}{\mathrm{d}\tau_{\mathrm{p}}}\frac{\mathrm{d}X^\mu}{\mathrm{d}\tau_{\mathrm{p}}}. \tag{15.1}$$

Here I denote a point on the particle's trajectory, or *worldline*, by $X^\mu(\tau_{\mathrm{p}})$ (and $X_\mu = \eta_{\mu\nu}X^\nu$), to distinguish it from the coordinates x^μ of a general spacetime point, and the proper time measured along this worldline by τ_{p}. This action, and its generalization to a curved spacetime, serve well enough to describe the motion of a classical particle, but there is a catch. The proper time τ_{p} is determined by the Minkowski line element $\mathrm{d}\tau_{\mathrm{p}}^2 = \eta_{\mu\nu}\,\mathrm{d}X^\mu \mathrm{d}X^\nu$, and this implies that the components of the 4-velocity $\mathrm{d}X^\mu/\mathrm{d}\tau_{\mathrm{p}}$ are not all independent, but are constrained by the relation

$$\frac{\mathrm{d}X_\mu}{\mathrm{d}\tau_{\mathrm{p}}}\frac{\mathrm{d}X^\mu}{\mathrm{d}\tau_{\mathrm{p}}} = 1. \tag{15.2}$$

So long as we deal with a classical particle, for which $X^\mu(\tau_{\mathrm{p}})$ is a definite, well-defined function, it is simple enough to add this equation to the Euler–Lagrange equations obtained from the action, as we did in (4.43), for example. For a quantum-mechanical particle, which does not have a well-defined worldline, matters are less straightforward. In fact, a large part of the mathematical complexity of string theory can be traced to the necessity of imposing a constraint similar to (15.2). For the point particle, one way of proceeding is to label points on the worldline by an arbitrary parameter τ. An element $\mathrm{d}\tau_{\mathrm{p}}$ of proper time must then be related to a small change in τ by $\mathrm{d}\tau_{\mathrm{p}} = e(\tau)\mathrm{d}\tau$, where $e(\tau)$ amounts to a metric on the worldline (or it might be thought of as analogous to the vierbein that we introduced in (7.130)). We then have $\mathrm{d}X^\mu/\mathrm{d}\tau_{\mathrm{p}} = e^{-1}\mathrm{d}X^\mu/\mathrm{d}\tau$, and a suitable action is

$$S = -\tfrac{1}{2}m\int \mathrm{d}\tau\left[e^{-1}\dot{X}_\mu\dot{X}^\mu + e\right] \tag{15.3}$$

where $\dot{X}^\mu = \mathrm{d}X^\mu/\mathrm{d}\tau$. Classically, we could choose $\tau = \tau_{\mathrm{p}}$ by setting $e = 1$, in which case this new action differs from (15.1) by an irrelevant constant. The point of introducing $e(\tau)$, though, is that we can treat it as a new dynamical variable, on the same footing as $X^\mu(\tau)$. If we do this, then the two Euler–Lagrange equations are

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{e}\frac{\mathrm{d}X^\mu}{\mathrm{d}\tau}\right)=0 \qquad \text{or} \qquad \frac{\mathrm{d}^2X^\mu}{\mathrm{d}\tau_\mathrm{p}^2}=0 \tag{15.4}$$

$$-e^{-2}\frac{\mathrm{d}X_\mu}{\mathrm{d}\tau}\frac{\mathrm{d}X^\mu}{\mathrm{d}\tau}+1=0 \qquad \text{or} \qquad \frac{\mathrm{d}X_\mu}{\mathrm{d}\tau_\mathrm{p}}\frac{\mathrm{d}X^\mu}{\mathrm{d}\tau_\mathrm{p}}=1. \tag{15.5}$$

Clearly, these two equations reproduce both the equation of motion for X^μ and the constraint (15.2). Equally clearly, we still have a theory that describes a single relativistic particle, so the function $e(\tau)$ cannot correspond to a genuine physical degree of freedom. In fact, it is a gauge degree of freedom, analogous to the component A_0 of the electromagnetic 4-vector potential which, as we saw in chapter 9, acts as a Lagrange multiplier to enforce the Gauss' law constraint (9.55). In the case at hand, the gauge symmetry is the freedom we have introduced to relabel points on the worldline. If we choose a new parameter $\tau'=\tau'(\tau)$, we must have $\mathrm{d}\tau_\mathrm{p}=e(\tau)\mathrm{d}\tau=e'(\tau')\mathrm{d}\tau'$ and it is a simple matter to check that the transformation

$$\mathrm{d}\tau=\frac{\mathrm{d}\tau}{\mathrm{d}\tau'}\mathrm{d}\tau' \qquad e=\frac{\mathrm{d}\tau'}{\mathrm{d}\tau}e' \qquad \frac{\mathrm{d}X^\mu}{\mathrm{d}\tau}=\frac{\mathrm{d}\tau'}{\mathrm{d}\tau}\frac{\mathrm{d}X'^\mu}{\mathrm{d}\tau'} \tag{15.6}$$

leaves the form of the action (15.3) unchanged. Clearly, this *reparametrization invariance* is quite analogous to the general-coordinate invariance of general relativity. Each spacetime coordinate X^μ counts as a scalar field for this purpose, which means, as in (2.9), that $X'^\mu(\tau')=X^\mu(\tau)$ when τ and τ' label the same point of the worldline.

I shall illustrate the quantum-mechanical use of this description of a relativistic particle by showing that the path integral

$$\Delta(x,y)=\mathcal{N}\int_x^y \mathcal{D}X(\tau)\mathcal{D}e(\tau)\mathrm{e}^{\mathrm{i}S} \tag{15.7}$$

where $\mathcal{N}$ is a normalizing constant, is the Feynman propagator (9.40), provided that we can find a suitable interpretation of the somewhat ill-defined integration measure $\mathcal{D}X(\tau)\mathcal{D}e(\tau)$. The limits on the integral indicate that it is a sum over worldlines that start at the spacetime point x and end at y. To be specific, we label points on the worldline by values of τ between 0 and 1, and impose the condition

$$y^\mu-x^\mu=X^\mu(1)-X^\mu(0)=\int_0^1\dot{X}^\mu(\tau)\mathrm{d}\tau \tag{15.8}$$

by inserting a δ function into the path integral. That is,

$$\Delta(x,y)=\mathcal{N}\int\mathcal{D}X(\tau)\mathcal{D}e(\tau)\exp(\mathrm{i}S)\,\delta^4\left(\textstyle\int_0^1\dot{X}\mathrm{d}\tau+x-y\right)$$

$$= \int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\mathcal{N}\int \mathcal{D}X(\tau)\mathcal{D}e(\tau)\exp\left(\mathrm{i}S - \mathrm{i}\textstyle\int_0^1 \mathrm{d}\tau\, k\cdot\dot{X}\right) \tag{15.9}$$

where the integration variables $X^\mu(\tau)$ include the endpoints. In the second expression I have used the representation of the δ function given in (A.11). The argument of the exponential can be written as $\mathrm{i}S_k$, where

$$S_k = -\frac{m}{2}\int_0^1 \mathrm{d}\tau\left[e^{-1}\left(\dot{X}_\mu + \frac{e}{m}k_\mu\right)\left(\dot{X}^\mu + \frac{e}{m}k^\mu\right) + e\left(1 - \frac{1}{m^2}k_\mu k^\mu\right)\right] \tag{15.10}$$

and after a change integration variable

$$X^\mu(\tau) \to X^\mu(\tau) - m^{-1}k^\mu \int_0^\tau \mathrm{d}\tau' e(\tau')$$

this becomes

$$S_k = -\frac{1}{2}\int_0^1 \mathrm{d}\tau\left[me^{-1}\dot{X}_\mu\dot{X}^\mu - m^{-1}e(k^2 - m^2)\right]. \tag{15.11}$$

Now, one part of the information contained in $e(\tau)$ is the total proper time along the particle's worldline,

$$\hat{\tau}_\mathrm{p} = \int_0^{\hat{\tau}_\mathrm{p}} \mathrm{d}\tau_\mathrm{p} = \int_0^1 \mathrm{d}\tau\, e(\tau) \tag{15.12}$$

so the path integral $\int \mathcal{D}e(\tau)$ includes an integral over all values of $\hat{\tau}_\mathrm{p}$. If $\widetilde{e}$ denotes the remaining degrees of freedom, then we have

$$\Delta(x,y) = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\int_0^\infty \mathrm{d}\hat{\tau}_\mathrm{p}\,\mathrm{e}^{\mathrm{i}(\hat{\tau}_\mathrm{p}/2m)(k^2-m^2)}$$
$$\times\,\mathcal{N}\int \mathcal{D}X\mathcal{D}\widetilde{e}\exp\left(-\tfrac{1}{2}\mathrm{i}m\int_0^1 \mathrm{d}\tau\, e^{-1}\dot{X}_\mu\dot{X}^\mu\right). \tag{15.13}$$

The remaining path integral is independent of k. Provided that the integration measure is appropriately defined, it is independent of $\hat{\tau}_\mathrm{p}$ too, so it is just a constant. If we choose $\mathcal{N}$ to be $(2m)^{-1}$ times this constant, define $\lambda = \hat{\tau}_\mathrm{p}/2m$ and, as in §9.3.2, introduce a convergence factor into the integral by changing m^2 into $m^2 - \mathrm{i}\epsilon$, we get

$$\Delta(x,y) = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}\int_0^\infty \mathrm{d}\lambda\,\mathrm{e}^{\mathrm{i}\lambda(k^2-m^2+\mathrm{i}\epsilon)}$$
$$= \mathrm{i}\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{\mathrm{e}^{-\mathrm{i}k\cdot(x-y)}}{k^2 - m^2 + \mathrm{i}\epsilon} \tag{15.14}$$

and this is indeed just i times the Feynman propagator.

The object we have computed is rather analogous to the generating functional (9.33) for a quantum field theory, but the field theory in question has fields $X^{\mu}(\tau)$, which are the particle's space-time coordinates, and it lives on a one-dimensional manifold, which is the particle's world line. Suppose that we take the same field theory and place it on a more complicated one-dimensional manifold, namely a Feynman diagram. It should appear plausible—and it may even be obvious—that a calculation analogous to the one we have just been through will yield the contribution of this diagram to the relevant scattering amplitude, as determined by rules (i)–(iii) in §9.4. The total scattering amplitude (as given, at least, by perturbation theory) is got by summing over all the allowed topologies of this one-dimensional manifold (a network of worldlines), and it is a generalization of this idea that constitutes the perturbative approach to string theory. A complete theory of point particles constructed in this way would be a rather *ad hoc* affair, for several reasons. One is that we would have to decide what topologies for the network of worldlines are allowed or, in other words, what vertices are allowed in rule (ii). Another is that we would have to insert by hand the coupling constants required by rule (ii) and the combinatorial factors required by rule (iv). A third is that we should have to find some way of generalizing the action (15.3) to account for the existence of particles of several different species, with different spins, and of specifying which parts of the worldline network are inhabited by which particle species. All of these matters are systematized in the second-quantized formalism of quantum field theory, where the theory is completely specified by the action for field operators living in spacetime. The lesson of chapter 12, though, is that we have no *a priori* way of knowing exactly what this action should be.

In string theory, we shall see that things are otherwise. The network of worldlines is replaced by a two-dimensional *worldsheet*. Although this worldsheet may have different topologies, which must be summed over, it has no well-defined vertices: there are no coupling constants or combinatorial factors to be specified. Different particle species correspond, in a way that I shall make more precise in the next section, to different modes of vibration of a single string-like object, so they all exist on the whole worldsheet. In this sense, string theory comes close to specifying a unique 'theory of everything'. There are, however, choices of a different kind to be made, about which we shall learn a little more later on, and the current theory is, essentially, only a perturbative one. Whether some overarching, nonperturbative definition of the theory, analogous to the definition of a quantum field theory of point particles, is possible, and whether this definition would be unique, is at present not clear.

It is worth observing that the role of spacetime is quite different in the first- and second-quantized theories of point particles. Quantum field theory, which we could notionally take to include a quantum theory of gravitons, is formulated in terms of field operators, which exist at each point of a pre-existing spacetime manifold. In the first-quantized theory, on the other hand, the field operators X^{μ} and e exist at each point of a different manifold, the network of

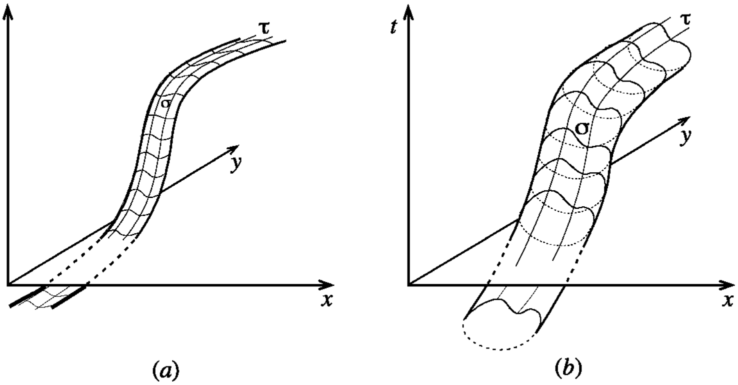


Figure 15.1. The worldsheet traced out by (a) an open string and (b) a closed string propagating through spacetime.

worldlines. Spacetime is just the set of values that the fields X^μ might take on: it is something that emerges from a more fundamental level of description. Whether these are simply two complementary points of view, one better adapted to each mathematical formalism than the other, or whether one is really more fundamental than the other is something I am not at all sure about.

15.2 The Free Classical String

Our study of string theory proper begins with the problem of finding a quantum-mechanical description of a one-dimensional object—a string—that propagates through Minkowski spacetime. In this section, I shall deal with the theory of a classical relativistic string, which we can subsequently attempt to quantize. In classical terms, then, this one-dimensional object traces out a two-dimensional worldsheet, which we can specify by giving the spacetime coordinates $X^\mu(\tau, \sigma)$ as functions of two coordinates τ and σ which label points on the worldsheet.

15.2.1 The string action

There are two obvious possibilities for the topology of our string: it might have two free ends, in which case it is said to be *open* and its worldsheet is a ribbon like that shown in figure 15.1(a); or it might form a closed loop, in which case it is said to be *closed* and its worldsheet is a cylindrical object such as that shown in figure 15.1(b). The coordinates τ and σ on the worldsheet are to a large extent arbitrary, but I shall always assume that a curve of constant σ, whose points are labelled by τ, runs along the length of the worldsheet. Regarded as a curve in spacetime, it has a timelike tangent vector. Conversely, a curve of constant τ, whose points are labelled by σ, has a spacelike tangent vector. On the ribbon-like worldsheet of an open string, it has one end-point on each of the two timelike boundaries; on the cylindrical worldsheet of a closed string it forms a closed loop, which runs once around the cylinder. While τ can take on values from $-\infty$ to ∞, the values of σ lie in a finite interval, which for the moment I shall take to be 0 to ℓ.

The action that has been found to work is

$$S = -\frac{1}{4\pi\alpha'}\int_{-\infty}^{\infty} \mathrm{d}\tau \int_0^{\ell} \mathrm{d}\sigma\, (-\gamma)^{1/2}\gamma^{ab}\partial_a X_\mu \partial_b X^\mu . \tag{15.15}$$

The indices a and b take the values 0 and 1 to label the worldsheet coordinates, with $\sigma^0 = \tau$ and $\sigma^1 = \sigma$. As for the point particle, we introduce a worldsheet metric γ_{ab} whose determinant is γ and whose inverse is γ^{ab}; the determinant is negative because the worldsheet has one timelike and one spacelike direction. (To say that this action has been found to work means that it is the starting point for what appears to be a mathematically consistent theory; whether this theory has anything to do with the real world is entirely a matter for speculation.) It should be clear from our earlier discussions of physics in curved spacetimes (see, in particular §§4.2 and 4.3) that the volume element $(-\gamma)^{1/2}\mathrm{d}\tau\mathrm{d}\sigma$ and the quantity $\gamma^{ab}\partial_a X_\mu \partial_b X^\mu$ both transform as scalars under worldsheet reparametrizations, and so S is reparametrization invariant. In fact, it is the two-dimensional version of the first term of the point-particle action (15.3). To see what has happened to the second term, consider the change of variable $\tau = m\tau'$ and take the limit $m \to 0$, as we did in §4.4.4 to find the path of a massless particle such as a photon. The first term remains intact, but the second vanishes. There is no such term in (15.15) because any one point of the string carries a mass of zero. By comparing (15.15) with (15.3), we might guess that the string has a mass per unit length (or *tension*—see §13.3) of $1/2\pi\alpha'$. That this is indeed so is illustrated in exercise 15.2, which readers may like to attempt after reading a little further. The constant that determines the string tension is conventionally denoted by α' for historical reasons that I propose not to discuss. Considered as a whole, the string carries a mass which is determined not only by its tension but also by the internal energy of its vibrations; how we can find out the mass of a vibrating quantum-mechanical string is a matter that will require careful attention.

Let us make some routine deductions from our action. The Euler–Lagrange equation obtained by varying $X^\mu(\tau, \sigma)$ is

$$\partial_a\left[(-\gamma)^{1/2}\gamma^{ab}\partial_b X^\mu\right] = 0 \qquad \text{or} \qquad \gamma^{ab}\nabla_a\nabla_b X^\mu = 0 \tag{15.16}$$

where ∇_a is the covariant derivative associated with the worldsheet metric γ_{ab}. The second version follows from the first because of the expression (A.22) for the divergence of a vector field. This Euler–Lagrange equation can be recognized as a two-dimensional version of the Klein–Gordon equation (7.129) in a curved spacetime. It is derived by the standard procedure that we met first in §3.1, but because the range of σ is finite, we must be careful about the boundary conditions. To be specific, the usual integration by parts gives us a boundary term

$$\delta S_{\text{boundary}} = -\frac{1}{2\pi\alpha'}\int \mathrm{d}\tau\, (-\gamma)^{1/2}\gamma^{1a}\partial_a X^\mu \delta X_\mu \Big|_{\sigma=0}^{\sigma=\ell} \tag{15.17}$$

and we need this to vanish. For a closed string, $\sigma = 0$ and $\sigma = \ell$ refer to the same point, so it does vanish identically. An open string has ends that are free to move, so we cannot assume that $\delta X_\mu = 0$. Instead, we must impose the boundary condition

$$\gamma^{1a}\partial_a X^\mu(\tau,\sigma) \equiv \partial^1 X^\mu = 0 \tag{15.18}$$

at $\sigma = 0$ and $\sigma = \ell$. This means that the derivative of X^μ is zero in the direction normal to the worldsheet boundary, as we can verify in the following way. Let $t^a = \delta_0^a$ be the components of the tangent vector $\boldsymbol{t} = \partial_\tau$ to the worldsheet boundary, and n^a the components of a vector normal to the boundary. The definition of 'normal' is provided by the metric γ_{ab}, so $n^a\gamma_{ab}t^b = n^a\gamma_{a0} = 0$. Using this, we can calculate the derivative of X^μ in the normal direction to be

$$n^a\partial_a X^\mu = n^a\gamma_{ab}\partial^b X^\mu = n^a\gamma_{a1}\partial^1 X^\mu = 0. \tag{15.19}$$

According to a conventional terminology in the theory of differential equations, the open string is said to satisfy *Neumann* boundary conditions.

The constraint equation that we get by varying the metric is

$$T^{ab}(\tau,\sigma) = 0 \tag{15.20}$$

where

$$T^{ab} = -4\pi(-\gamma)^{-1/2}\frac{\delta S}{\delta\gamma_{ab}} = -\frac{1}{\alpha'}\left[\partial^a X_\mu \partial^b X^\mu - \frac{1}{2}\gamma^{ab}\partial_c X_\mu \partial^c X^\mu\right] \tag{15.21}$$

is the energy–momentum tensor of the worldsheet field theory. [To be clear about the notation here, ∂^a is an abbreviation for $\gamma^{ab}\partial_b = \gamma^{ab}\partial/\partial\sigma^b$. Below, I shall use ∂_τ and ∂_σ to mean the same thing as ∂_0 and ∂_1, respectively.] This is in fact the two-dimensional version of Einstein's field equations (4.17) with $\Lambda = 0$, because the Einstein curvature tensor $R_{ab} - \frac{1}{2}R\gamma_{ab}$ vanishes identically in two dimensions (see exercise 15.1). The energy–momentum tensor will play a central role in the development of the theory, and we may note at this point that it obeys the equation

$$\nabla_a T^{ab} = 0 \tag{15.22}$$

regardless of the constraint (15.20), which we also want to impose. This equation is also true in general relativity, although I have not needed to emphasize it. Here, it represents the conservation of energy and momentum flowing on the worldsheet. It is a consequence of the reparametrization invariance of the action (which is also commonly referred to as *diffeomorphism invariance*) and can be derived from a suitable version of Noether's theorem. Readers should find it a simple matter, though, to verify (15.22) directly from the equation of motion (15.16), bearing in mind that $\partial_a X^\mu = \nabla_a X^\mu$, because X^μ is a scalar field on the worldsheet, and that $\nabla_a \gamma_{bc} = 0$ (see §2.3.5).

As always, we can find a momentum $\Pi^\mu(\tau, \sigma)$ conjugate to the field $X_\mu(\tau, \sigma)$. Taking account of the sign in (3.33) arising from the Minkowski metric, we find

$$\Pi^\mu(\tau, \sigma) = -\frac{\delta S}{\delta \dot{X}_\mu(\tau, \sigma)} = \frac{1}{2\pi\alpha'}(-\gamma)^{1/2}\gamma^{0a}\partial_a X^\mu(\tau, \sigma) \qquad (15.23)$$

where $\dot{X}_\mu = \partial_\tau X_\mu$. The action (15.15) is obviously invariant under spacetime translations $X^\mu \to X^\mu + a^\mu$, because it depends only on derivatives of X^μ. The version of Noether's theorem given in (3.12) applies here if we substitute $\int \mathrm{d}\sigma$ for $\sum_i$, so we learn that the quantities

$$P^\mu = \int_0^\ell \mathrm{d}\sigma\, \Pi^\mu(\tau, \sigma) = \frac{1}{2\pi\alpha'}\int_0^\ell \mathrm{d}\sigma\, (-\gamma)^{1/2}\gamma^{0a}\partial_a X^\mu(\tau, \sigma) \qquad (15.24)$$

are conserved, in the sense that $\partial_\tau P^\mu = 0$. This can also be verified by using the equation of motion (15.16) and, in the case of an open string, the boundary conditions (15.18). The fact that the P^μ are independent of τ means that they are constant along the length of the worldsheet, so they can be identified as the components of the conserved spacetime momentum carried by the string. In the same way, the generators of Lorentz transformations in spacetime are

$$M^{\mu\nu} = \int_0^\ell \mathrm{d}\sigma \left[X^\mu(\tau, \sigma)\Pi^\nu(\tau, \sigma) - X^\nu(\tau, \sigma)\Pi^\mu(\tau, \sigma)\right] \qquad (15.25)$$

and from these we can identify the angular momentum $J^i = \frac{1}{2}\epsilon^{ijk}M^{jk}$ as in (7.42).

15.2.2 Weyl invariance and gauge fixing

In addition to diffeomorphism invariance, the action (15.15) has a further symmetry, which will prove important. Consider the effect of changing the worldsheet metric by a position-dependent factor

$$\gamma'_{ab}(\tau, \sigma) = \exp[\omega(\tau, \sigma)]\gamma_{ab}(\tau, \sigma) \qquad (15.26)$$

where $\omega(\tau, \sigma)$ is an arbitrary function (except that on a closed worldsheet it must be periodic, so that $\omega(\tau, \sigma + \ell) = \omega(\tau, \sigma)$); we use the exponential to ensure that the sign of the metric does not change. This rescaling of the metric is called a *Weyl transformation*. The determinant γ changes by a factor of $\exp(2\omega)$ and the inverse metric γ^{ab} changes by a factor of $\exp(-\omega)$, so the action is unchanged. (This symmetry is equivalent to one that I mentioned in chapter 7 under the name of 'conformal invariance'. In the context of string theory, the term 'conformal invariance' is used in a different, though closely related sense, which we shall meet before long.) An immediate consequence of this symmetry is that

$$\frac{\delta S}{\delta\omega(\tau,\sigma)} = \frac{\partial\gamma_{ab}(\tau,\sigma)}{\partial\omega(\tau,\sigma)}\frac{\delta S}{\delta\gamma_{ab}(\tau,\sigma)} = 0 \tag{15.27}$$

or, according to the definition (15.21) of the energy–momentum tensor,

$$\gamma_{ab}T^{ab} = T^a_a = 0. \tag{15.28}$$

It is easy to check that this is true of the explicit expression given in (15.21).

The combined symmetries of diffeomorphism invariance and Weyl invariance constitute a gauge symmetry of the string action involving three arbitrary functions, namely $\omega(\tau,\sigma)$ and the two functions $\tau'(\tau,\sigma)$ and $\sigma'(\tau,\sigma)$ which define a change of coordinates. It is a special feature of two-dimensional geometry that the metric has three independent components, *viz.* γ_{00}, γ_{11} and $\gamma_{01} = \gamma_{10}$. By using coordinate and Weyl transformations, it is possible to bring the worldsheet metric into the form $\gamma_{ab} = \eta_{ab}$, where η_{ab} is the two-dimensional version of the Minkowski metric (2.8), with diagonal components $\eta_{00} = 1$ and $\eta_{11} = -1$. In fact, it is possible to show (although a detailed proof is not entirely straightforward) that given any two-dimensional metric with one positive and one negative eigenvalue, a coordinate system can always be found in which the metric tensor has the form

$$\gamma_{ab}(\tau,\sigma) = \exp[\Omega(\tau,\sigma)]\eta_{ab}. \tag{15.29}$$

A Weyl transformation with $\omega = -\Omega$ then reduces the metric to just η_{ab}.

As far as classical mechanics is concerned, the physical content of the point-particle theory is contained in the second versions of (15.4) and (15.5), which can be solved to find the allowed worldlines, parametrized by the proper time τ_{p}. The function $e(\tau)$ has no physical meaning, and we are perfectly entitled to 'fix the gauge' by choosing any function we like, bearing in mind that our choice also implies a choice of the coordinate τ, such that the proper time is given by $\mathrm{d}\tau_{\mathrm{p}} = e(\tau)\mathrm{d}\tau$. Obviously, the most convenient choice is $e = 1$ and $\tau = \tau_{\mathrm{p}}$. For the classical string, we are equally entitled to fix the gauge by making use of the diffeomorphism and Weyl symmetries to choose $\gamma_{ab}(\tau,\sigma) = \eta_{ab}$. This does not uniquely specify a pair of worldsheet coordinates, though, because a 2-dimensional Lorentz transformation of these coordinates leaves the metric η_{ab} unchanged. To see that the physical content of the theory is independent of this gauge choice, suppose first that we have identified an allowed worldsheet by solving (15.16) subject to the constraint (15.20). Then the proper time along any curve drawn on this worldsheet is given by

$$\mathrm{d}\tau_{\mathrm{p}} = \eta_{\mu\nu}\mathrm{d}X^\mu\mathrm{d}X^\nu = \frac{\partial X_\mu}{\partial\sigma^a}\frac{\partial X^\mu}{\partial\sigma^b}\mathrm{d}\sigma^a\mathrm{d}\sigma^b \tag{15.30}$$

because the infinitesimal difference in the spacetime coordinates of two points at σ^a and $\sigma^a + \mathrm{d}\sigma^a$ on the worldsheet is given by $\mathrm{d}X^\mu = \partial_a X^\mu \mathrm{d}\sigma^a$. This proper time is clearly invariant under a change in the worldsheet coordinates σ^a. As for the equations (15.16) and (15.20) themselves, they are covariant under transformations of the worldsheet coordinates. Under a Weyl transformation, they become a different pair of equations, but because the action is invariant, its extrema, which are the allowed worldsheets, can be found by solving either pair of equations.

Let us, then, choose $\gamma_{ab} = \eta_{ab}$. With this choice, the content of the theory as we have it so far is summarized by

the action:

$$S = -\frac{1}{4\pi\alpha'}\int \mathrm{d}\tau\mathrm{d}\sigma\ \partial_a X_\mu \partial^a X^\mu \tag{15.31}$$

the canonical momentum:

$$\Pi^\mu(\tau,\sigma) = \frac{1}{2\pi\alpha'}\,\partial_\tau X^\mu(\tau,\sigma) \tag{15.32}$$

the spacetime momentum:

$$P^\mu = \int_0^\ell \mathrm{d}\sigma\ \Pi^\mu(\tau,\sigma) = \frac{1}{2\pi\alpha'}\int_0^\ell \mathrm{d}\sigma\ \partial_\tau X^\mu(\tau,\sigma) \tag{15.33}$$

the energy–momentum tensor:

$$T^{ab} = -\frac{1}{\alpha'}\left[\partial^a X_\mu \partial^b X^\mu - \frac{1}{2}\eta^{ab}\partial_c X_\mu \partial^c X^\mu\right] \tag{15.34}$$

the equation of motion:

$$\left[\partial_\tau^2 - \partial_\sigma^2\right] X^\mu = 0 \tag{15.35}$$

the constraint:

$$T^{ab} = 0 \tag{15.36}$$

energy–momentum conservation:

$$\partial_a T^{ab} = 0. \tag{15.37}$$

It is perhaps worth emphasizing that we now have two independent metrics on the worldsheet, which is to say that there are two different definitions of the ‘length’ of a curve drawn on it. If such a curve is thought of as a curve in spacetime, then its length depends on the values of the $X^\mu(\tau,\sigma)$, which determine how the worldsheet is embedded in spacetime, and is given by the line element (15.30). Classically, this length has an unambiguous physical meaning, but we shall actually not be making much use of it. For the purpose of dealing with the two-dimensional field theory of the X^μ, the manifold on which these fields live has the gauge-fixed metric η_{ab}, and the length of a curve is determined by the line element $\mathrm{d}\tau_{\mathrm{ws}}^2 = \mathrm{d}\tau^2 - \mathrm{d}\sigma^2$. The proper time interval $\mathrm{d}\tau_{\mathrm{p}}$ apparent to an observer in spacetime is in general quite different from the proper time interval $\mathrm{d}\tau_{\mathrm{ws}}$ defined on the worldsheet.

15.2.3 The Euclidean worldsheet and conformal invariance

A mathematical device that turns out to be useful is the Wick rotation, which we discussed in connection with (10.102). Here, we replace the Minkowskian metric on the worldsheet with a Euclidean one, by making the change of variable $\tau = -\mathrm{i}\sigma^2$. This is particularly helpful in two dimensions, because we can make use of complex variable theory by defining the single complex coordinate w and its complex conjugate $\bar{w}$ as

$$w = \sigma^1 + \mathrm{i}\sigma^2 = \sigma - \tau \qquad \bar{w} = \sigma^1 - \mathrm{i}\sigma^2 = \sigma + \tau. \tag{15.38}$$

In terms of w and $\bar{w}$, the coordinates τ and σ are

$$\tau = -\tfrac{1}{2}(w - \bar{w}) \qquad \sigma = \tfrac{1}{2}(w + \bar{w}). \tag{15.39}$$

It becomes a little inconvenient to label the components of tensors relative to the $(w, \bar{w})$ coordinates by numerical indices. For derivatives, the conventional notation is

$$\partial \equiv \frac{\partial}{\partial w} = \frac{1}{2}\left(\frac{\partial}{\partial\sigma} - \frac{\partial}{\partial\tau}\right) \qquad \bar{\partial} \equiv \frac{\partial}{\partial\bar{w}} = \frac{1}{2}\left(\frac{\partial}{\partial\sigma} + \frac{\partial}{\partial\tau}\right). \tag{15.40}$$

A vector with components $V^\tau \equiv V^0$ and $V^\sigma \equiv V^1$ relative to the (τ, σ) system has components V^w and $V^{\bar{w}}$ relative to the $(w, \bar{w})$ system, which are given by

$$\begin{pmatrix} V^w \\ V^{\bar{w}} \end{pmatrix} = \Lambda \begin{pmatrix} V^\tau \\ V^\sigma \end{pmatrix} \tag{15.41}$$

where the matrix Λ defined in (2.13) is

$$\Lambda = \begin{pmatrix} \partial w/\partial\tau & \partial w/\partial\sigma \\ \partial\bar{w}/\partial\tau & \partial\bar{w}/\partial\sigma \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \tag{15.42}$$

and its inverse, which transforms covariant indices, is

$$\Lambda^{-1} = \begin{pmatrix} \partial\tau/\partial w & \partial\tau/\partial\bar{w} \\ \partial\sigma/\partial w & \partial\sigma/\partial\bar{w} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{15.43}$$

Thus, the gauge-fixed metric has components

$$\begin{pmatrix} \gamma_{ww} & \gamma_{w\bar{w}} \\ \gamma_{\bar{w}w} & \gamma_{\bar{w}\bar{w}} \end{pmatrix} = (\Lambda^{-1})^{\mathrm{T}} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \Lambda^{-1} = -\frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{15.44}$$

which means that a proper *distance* on the worldsheet (as specified by the worldsheet metric, not by the spacetime metric) is

$$\mathrm{d}s^2 = -\mathrm{d}\tau_{\mathrm{ws}}^2 = -\mathrm{d}\tau^2 + \mathrm{d}\sigma^2 = (\mathrm{d}\sigma^1)^2 + (\mathrm{d}\sigma^2)^2 = \mathrm{d}w\,\mathrm{d}\bar{w}. \tag{15.45}$$

In the same way, we find that the energy–momentum tensor is

$$\begin{pmatrix} T_{ww} & T_{w\bar{w}} \\ T_{\bar{w}w} & T_{\bar{w}\bar{w}} \end{pmatrix} = (\Lambda^{-1})^{\mathrm{T}} \begin{pmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{pmatrix} \Lambda^{-1} = \begin{pmatrix} T & 0 \\ 0 & \widetilde{T} \end{pmatrix} \tag{15.46}$$

where the nonzero components are

$$T = -\frac{1}{\alpha'}\partial X_\mu \partial X^\mu \qquad \widetilde{T} = -\frac{1}{\alpha'}\bar{\partial} X_\mu \bar{\partial} X^\mu. \tag{15.47}$$

As in (10.102), we define an action on the Euclidean worldsheet by $\mathrm{i}S = -S_{\mathrm{E}}$. To get the right answer for S_{E}, starting from (15.31), we must be careful to treat the volume element $\mathrm{d}\tau\mathrm{d}\sigma$ correctly. First, we make an analytic continuation from real to imaginary time, which means replacing $\mathrm{d}\tau\mathrm{d}\sigma$ with $-\mathrm{i}\mathrm{d}\sigma^1\mathrm{d}\sigma^2$. Thereafter, the change of variables from (σ^1, σ^2) to $(w, \bar{w})$ yields a Jacobian, which is $|\partial(\sigma^1, \sigma^2)/\partial(w, \bar{w})| = \frac{1}{2}$. The result is

$$\begin{aligned} S_{\mathrm{E}} &= -\frac{1}{4\pi\alpha'}\int \mathrm{d}^2\sigma \left[\partial_1 X_\mu \partial_1 X^\mu + \partial_2 X_\mu \partial_2 X^\mu\right] \\ &= -\frac{1}{2\pi\alpha'}\int \mathrm{d}w\mathrm{d}\bar{w}\, \partial X_\mu \bar{\partial} X^\mu. \end{aligned} \tag{15.48}$$

The field theory defined by this action has a crucial symmetry, known as *conformal invariance*. It will perhaps be helpful to discuss this symmetry from two complementary points of view. Consider first the idea of replacing the fields $X^\mu(w, \bar{w})$ by a new set of fields

$$X'^\mu(w, \bar{w}) = X^\mu(f(w), \bar{f}(\bar{w})) \tag{15.49}$$

where $f(w)$ is an arbitrary function of w, but is independent of $\bar{w}$, and $\bar{f}(\bar{w})$ is the complex conjugate of $f(w)$. We have

$$\partial X'_\mu \bar{\partial} X'^\mu = \frac{\mathrm{d}f}{\mathrm{d}w}\frac{\mathrm{d}\bar{f}}{\mathrm{d}\bar{w}}\frac{\partial X_\mu}{\partial f}\frac{\partial X^\mu}{\partial \bar{f}}. \tag{15.50}$$

To find the action of the new fields, we introduce new integration variables

$$w' = f(w) \qquad \bar{w}' = \bar{f}(\bar{w}) \tag{15.51}$$

and calculate

$$\begin{aligned} S' &= -\frac{1}{2\pi\alpha'}\int \mathrm{d}w\mathrm{d}\bar{w}\frac{\partial X'_\mu}{\partial w}\frac{\partial X'^\mu}{\partial \bar{w}} \\ &= -\frac{1}{2\pi\alpha'}\int \mathrm{d}w'\mathrm{d}\bar{w}'\frac{\mathrm{d}w}{\mathrm{d}w'}\frac{\mathrm{d}\bar{w}}{\mathrm{d}\bar{w}'}.\frac{\mathrm{d}w'}{\mathrm{d}w}\frac{\mathrm{d}\bar{w}'}{\mathrm{d}\bar{w}}\frac{\partial X_\mu}{\partial w'}\frac{\partial X^\mu}{\partial \bar{w}'} \\ &= -\frac{1}{2\pi\alpha'}\int \mathrm{d}w'\mathrm{d}\bar{w}'\frac{\partial X_\mu}{\partial w'}\frac{\partial X^\mu}{\partial \bar{w}'}. \end{aligned} \tag{15.52}$$

The last expression is equal to S, because w' and $\bar{w}'$ are dummy integration variables, which we can replace with w and $\bar{w}$. Thus, the action calculated with the new fields $X'^{\mu}(w)$ is equal to that calculated with the old fields $X^{\mu}(w)$ and the transformation (15.49) is a symmetry of the theory. From another point of view, the change of variables (15.51), which is called a *conformal transformation* in the theory of complex variables, looks suspiciously like a simple change of coordinates on the worldsheet, so it is tempting to think that conformal invariance is just our original diffeomorphism invariance under another name. This is not quite true, because our gauge-fixed action is supposed to describe a field theory on a Euclidean worldsheet whose metric is given by the line element (15.45). In terms of the new coordinates (15.51), the line element is

$$\mathrm{d}s^2 = \mathrm{d}w\mathrm{d}\bar{w} = \mathrm{e}^{\Omega}\mathrm{d}w'\mathrm{d}\bar{w}' \qquad \Omega = -\ln\left|\frac{\mathrm{d}f}{\mathrm{d}w}\right|^2. \tag{15.53}$$

To make the theories described by S and S' completely equivalent, we have to remove the factor of e^{Ω} by making a Weyl transformation. The second view of conformal invariance, then, is that it constitutes a special combination of diffeomorphism and Weyl transformations. It is a remnant of the original gauge symmetry that is not removed by our choice of the metric.

As a prelude to examining the quantum-mechanical status of conformal invariance, it will be useful to identify the generators of this symmetry, which turn out to be the components T and $\widetilde{T}$ of the energy–momentum tensor. Classically, we need to discover how the field transformations (15.49) can be generated by Poisson brackets, as we did in §3.4 for spacetime translations. Here, we have an infinite number of generalized coordinates, namely the fields $X^{\mu}(\tau,\sigma)$ for every value of σ, and a suitable definition of the equal-τ Poisson bracket of two quantities $A(\tau)$ and $B(\tau)$ is

$$\{A(\tau), B(\tau)\}_{\mathrm{P}} = -\int_0^{\ell}\mathrm{d}\sigma\left[\frac{\delta A(\tau)}{\delta X^{\mu}(\tau,\sigma)}\frac{\delta B(\tau)}{\delta \Pi_{\mu}(\tau,\sigma)} - \frac{\delta B(\tau)}{\delta X^{\mu}(\tau,\sigma)}\frac{\delta A(\tau)}{\delta \Pi_{\mu}(\tau,\sigma)}\right]. \tag{15.54}$$

By expressing T and $\widetilde{T}$ in terms of $\partial_{\sigma}X^{\mu}$ and Π_{μ}, readers should have little trouble in verifying that

$$\begin{aligned}
\{X^{\mu}(\tau,\sigma), \Pi^{\nu}(\tau,\sigma')\}_{\mathrm{P}} &= -\eta^{\mu\nu}\delta(\sigma-\sigma') && (15.55)\\
\{X^{\mu}(\tau,\sigma), T(\tau,\sigma')\}_{\mathrm{P}} &= \tfrac{1}{2}\left[\partial_{\tau}X^{\mu}(\tau,\sigma) - \partial_{\sigma}X^{\mu}(\tau,\sigma)\right]2\pi\delta(\sigma-\sigma') \\
&= -2\pi\delta(\sigma-\sigma')\partial X^{\mu}(w,\bar{w}) && (15.56)\\
\{X^{\mu}(\tau,\sigma), \widetilde{T}(\tau,\sigma')\}_{\mathrm{P}} &= \tfrac{1}{2}\left[\partial_{\tau}X^{\mu}(\tau,\sigma) + \partial_{\sigma}X^{\mu}(\tau,\sigma)\right]2\pi\delta(\sigma-\sigma') \\
&= 2\pi\delta(\sigma-\sigma')\bar{\partial}X^{\mu}(w,\bar{w}). && (15.57)
\end{aligned}$$

The overall sign in (15.54) is determined by the fact that the *spatial* coordinates and momenta X^i and Π^j have a Poisson bracket $-\eta^{ij}\delta(\sigma-\sigma') = +\delta^{ij}\delta(\sigma-\sigma')$ with the same sign as their non-relativistic counterparts in chapter 3.

Consider now an infinitesimal version of the conformal transformation (15.49), in which we take $f(w) = w + \epsilon(w)$ and keep only first-order terms in $\epsilon(w)$. The infinitesimal change $\delta X^\mu = X'^\mu - X^\mu$ is

$$\begin{aligned}\delta X^\mu(w, \bar{w}) &= \epsilon(w)\partial X^\mu(w, \bar{w}) + \bar{\epsilon}(\bar{w})\bar{\partial} X^\mu(w, \bar{w}) \\ &= \{\boldsymbol{T}(\epsilon, \bar{\epsilon}), X^\mu(w, \bar{w})\}_{\mathrm{P}} \end{aligned} \qquad (15.58)$$

where

$$\boldsymbol{T}(\epsilon, \bar{\epsilon}) = \int_0^\ell \frac{\mathrm{d}\sigma'}{2\pi} \left[\epsilon(\sigma' - \tau)T(\tau, \sigma') - \bar{\epsilon}(\sigma' + \tau)\widetilde{T}(\tau, \sigma')\right]. \qquad (15.59)$$

We shall see later that this can be more neatly expressed as a contour integral when X^μ is a solution of the equation of motion. Note that $\boldsymbol{T}$ is not itself the generator of conformal transformations, because it contains the small parameters ϵ and $\bar{\epsilon}$. There are in fact infinitely many small parameters, namely the infinitely many functions $\epsilon(w)$. Correspondingly, there are infinitely many generators, all of which are contained in the integral (15.59).

These infinitely many generators (which I shall shortly be discussing in more detail) constitute the Lie algebra of the *conformal group*, and they confer a rich structure on a two-dimensional field theory that is conformally invariant. There is, in fact, a branch of theoretical physics, known as *conformal field theory*, which studies the consequences of conformal invariance in a rather general way. The mathematical techniques of conformal field theory are extremely valuable to professional string theorists, but I do not have space to develop them here. In statistical mechanics, the same techniques have had a remarkably unifying effect on the study of phase transitions in a large class of theoretical models, at which I hinted in §13.3. Readers who wish to pursue these ideas will find conformal field theory developed in the context of string theory by Polchinski (1998) and in the context of statistical mechanics by Cardy (1987).

15.2.4 Mode expansions

The gauge-fixed equation of motion (15.35) is known to every first-year undergraduate as the one-dimensional wave equation. Its general solution is the sum of an arbitrary function of $\sigma - \tau$ (a 'right-moving' wave) and an arbitrary function of $\sigma + \tau$ (a 'left-moving' wave). For a string of finite length ℓ, this general solution can be expressed as a Fourier series or, as is often said, a *mode expansion*. Let us recall that the value of ℓ is entirely arbitrary: it determines only the range of the coordinate σ, and not the actual length of the string. In the case of a closed string, it is now convenient to choose $\ell = 2\pi$, so that X^μ is a periodic function of σ, with $X^\mu(\tau, \sigma + 2\pi) = X^\mu(\tau, \sigma)$. The solution for the closed string is then

$$X^\mu(\tau,\sigma) = x^\mu + \alpha' p^\mu \tau + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{n}\left[\alpha_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} + \widetilde{\alpha}_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau+\sigma)}\right] \tag{15.60}$$

where, to make X^μ real, the expansion coefficients for positive and negative values of n must be related by

$$\alpha_n^{\mu *} = \alpha_{-n}^\mu \qquad \widetilde{\alpha}_n^{\mu *} = \widetilde{\alpha}_{-n}^\mu . \tag{15.61}$$

(The * here means the complex conjugate. At the classical level, it is really the same as the ¯ that distinguishes the two complex coordinates w and $\bar{w}$, but it is useful to have a different notation for this geometrical meaning. In the quantum theory, we shall want to replace $\alpha_n^{\mu *}$ with the Hermitian conjugate $\alpha_n^{\mu \dagger}$, but this would not make sense for the coordinates.) The first two terms in (15.60) are, of course, the sum of a function of $\sigma - \tau$ and a function of $\sigma + \tau$, namely $\frac{1}{2}[x^\mu + \alpha' p^\mu(\tau \pm \sigma)]$. By integrating $X^\mu(\tau,\sigma)$ over σ, we find

$$\int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi} X^\mu(\tau,\sigma) = x^\mu + \alpha' p^\mu \tau \tag{15.62}$$

which might loosely be thought of as the locating the centre of mass of the string, although the curve on the worldsheet that we are integrating over, corresponding to a fixed value of τ, does not necessarily represent an instantaneous configuration of the string as seen by some inertial observer in spacetime. We see from (15.33), however, that p^μ is equal to the spacetime momentum P^μ and it is a simple exercise using the Poisson bracket (15.55) to verify that x^μ and p^μ are conjugate variables, in the sense that

$$\{x^\mu, p^\nu\}_{\mathrm{P}} = -\eta^{\mu\nu}. \tag{15.63}$$

In a similar way, we can find expressions for the coefficients α_n^μ and $\widetilde{\alpha}_n^\mu$ analogous to (7.12) and (7.13). The mode expansion of the canonical momentum, found by differentiating (15.60), is

$$(2\pi\alpha')\Pi^\mu(\tau,\sigma) = \alpha' p^\mu + \left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \left[\alpha_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} + \widetilde{\alpha}_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau+\sigma)}\right] \tag{15.64}$$

and readers may easily verify, using the orthogonality relation

$$\int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi} \mathrm{e}^{\pm \mathrm{i}n\sigma} = \delta_{n,0} \tag{15.65}$$

that the expansion coefficients are given by

$$\alpha_n^\mu = \left(\frac{1}{2\alpha'}\right)^{1/2} \int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi}\, \mathrm{e}^{\mathrm{i}n(\tau-\sigma)} \left[(2\pi\alpha')\Pi^\mu(\tau,\sigma) - \mathrm{i}nX^\mu(\tau,\sigma)\right] \quad (15.66)$$

$$\widetilde{\alpha}_n^\mu = \left(\frac{1}{2\alpha'}\right)^{1/2} \int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi}\, \mathrm{e}^{\mathrm{i}n(\tau+\sigma)} \left[(2\pi\alpha')\Pi^\mu(\tau,\sigma) - \mathrm{i}nX^\mu(\tau,\sigma)\right]. \quad (15.67)$$

Their Poisson bracket relations

$$\{\alpha_m^\mu, \alpha_n^\nu\}_\mathrm{P} = \{\widetilde{\alpha}_m^\mu, \widetilde{\alpha}_n^\nu\}_\mathrm{P} = \mathrm{i}m\eta^{\mu\nu}\delta_{m,-n} \qquad \{\alpha_m^\mu, \widetilde{\alpha}_n^\nu\}_\mathrm{P} = 0. \quad (15.68)$$

follow straightforwardly from (15.55). We can also see from (15.64) that it consistent to define

$$\alpha_0^\mu = \widetilde{\alpha}_0^\mu = \left(\frac{1}{2\alpha'}\right)^{1/2} \int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi}\, (2\pi\alpha')\Pi^\mu(\tau,\sigma) = (\alpha'/2)^{1/2} p^\mu \quad (15.69)$$

which is useful for dealing with the derivatives of X^μ, although it cannot be used directly in (15.60) because of the factor of $1/n$.

One part of the task that confronts us in quantizing the theory is familiar from chapter 7, namely to promote the coefficients α_n^μ and $\widetilde{\alpha}_n^\mu$ to operators and their Poisson brackets to commutators. We shall find that there is more to it than that, however, and we need one more piece of classical theory to equip us, namely the algebra of the conformal generators. To attack this, we first express the mode expansion (15.60) in terms of the complex coordinates w and $\bar{w}$ as

$$X^\mu(w,\bar{w}) = x^\mu + \tfrac{1}{2}\alpha' p^\mu(\bar{w} - w) + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{n}\left[\alpha_n^\mu \mathrm{e}^{\mathrm{i}nw} + \widetilde{\alpha}_n^\mu \mathrm{e}^{-\mathrm{i}n\bar{w}}\right]. \quad (15.70)$$

It is evidently the sum of a function of w and a function of $\bar{w}$ and we can conclude that ∂X^μ is a function only of w while $\bar{\partial} X^\mu$ is a function only of $\bar{w}$. In the language of complex-variable theory, we say that ∂X^μ is *holomorphic* and $\bar{\partial} X^\mu$ is *antiholomorphic*. When X^μ is a solution of the field equations, therefore, equations (15.47) show us that $T = T(w)$ is holomorphic and $\widetilde{T} = \widetilde{T}(\bar{w})$ is antiholomorphic. In fact, by using the definition (15.69), we can write

$$\partial X^\mu = -\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{n=-\infty}^{\infty} \alpha_n^\mu \mathrm{e}^{\mathrm{i}nw} \qquad \bar{\partial} X^\mu = \left(\frac{\alpha'}{2}\right)^{1/2} \sum_{n=-\infty}^{\infty} \widetilde{\alpha}_n^\mu \mathrm{e}^{-\mathrm{i}n\bar{w}} \quad (15.71)$$

where the sums now include the terms $n = 0$. A short calculation reveals that $T(w)$ and $\widetilde{T}(\bar{w})$ can be expressed as

$$T(w) = \sum_{n=-\infty}^{\infty} L_n \mathrm{e}^{\mathrm{i}nw} \qquad \widetilde{T}(\bar{w}) = \sum_{n=-\infty}^{\infty} \widetilde{L}_n \mathrm{e}^{-\mathrm{i}n\bar{w}} \quad (15.72)$$

with the coefficients given by

$$L_n = -\tfrac{1}{2}\sum_{m=-\infty}^{\infty} \alpha_{m\,\mu}\alpha^{\mu}_{n-m} \qquad \widetilde{L}_n = -\tfrac{1}{2}\sum_{m=-\infty}^{\infty} \widetilde{\alpha}_{m\,\mu}\widetilde{\alpha}^{\mu}_{n-m}. \tag{15.73}$$

These coefficients are the generators of conformal transformations, and we need to know the Poisson-bracket relations between them. For two quantities A and B that depend only on the α_n^μ (and not on x^μ or $\widetilde{\alpha}_n^\mu$), we can use the functional derivatives of (15.66)

$$\frac{\delta\alpha_n^\mu}{\delta X^\nu(\tau,\sigma)} = -\frac{\mathrm{i}n}{2\pi}\left(\frac{1}{2\alpha'}\right)^{1/2}\delta^\mu_\nu \mathrm{e}^{\mathrm{i}n(\tau-\sigma)} \tag{15.74}$$

$$\frac{\delta\alpha_n^\mu}{\delta \Pi_\nu(\tau,\sigma)} = \left(\frac{\alpha'}{2}\right)^{1/2}\eta^{\mu\nu} \mathrm{e}^{\mathrm{i}n(\tau-\sigma)} \tag{15.75}$$

and the orthogonality relation (15.65) to express the Poisson bracket (15.54) as

$$\begin{aligned}\{A,B\}_{\mathrm{P}} &= -\int_0^{2\pi}\mathrm{d}\sigma\sum_{m,n}\left[\frac{\partial A}{\partial\alpha_m^\nu}\frac{\partial B}{\partial\alpha_n^\lambda} - \frac{\partial B}{\partial\alpha_m^\nu}\frac{\partial A}{\partial\alpha_n^\lambda}\right]\frac{\delta\alpha_m^\nu}{\delta X^\mu(\tau,\sigma)}\frac{\delta\alpha_n^\lambda}{\delta\Pi_\mu(\tau,\sigma)}\\ &= \frac{\mathrm{i}}{2}\sum_m m\left[\frac{\partial A}{\partial\alpha_{m\,\nu}}\frac{\partial B}{\partial\alpha^\nu_{-m}} - \frac{\partial B}{\partial\alpha_{m\,\nu}}\frac{\partial A}{\partial\alpha^\nu_{-m}}\right].\end{aligned} \tag{15.76}$$

Applying this to L_m and L_n, and using the same method for $\widetilde{L}_m$ and $\widetilde{L}_n$, we obtain (see exercise 15.3) the Poisson bracket relations

$$\{L_m, L_n\}_{\mathrm{P}} = -\mathrm{i}(m-n)L_{m+n} \qquad \{\widetilde{L}_m,\widetilde{L}_n\}_{\mathrm{P}} = -\mathrm{i}(m-n)\widetilde{L}_{m+n}. \tag{15.77}$$

A calculation similar to (15.76) shows that $\{A,B\}_{\mathrm{P}} = 0$ if A depends only on the α_n^μ while B depends only on the $\widetilde{\alpha}_n^\mu$ so we also have $\{L_m,\widetilde{L}_n\}_{\mathrm{P}} = 0$. The set of generators obeying these relations is called (after its discoverer, M A Virasoro) the *Virasoro algebra*.

A definite state of motion of our classical string would be specified by giving the values of x^μ and p^μ which, roughly speaking, describe the motion of its centre of mass, and the values of the α_n^μ and $\widetilde{\alpha}_n^\mu$, which are the amplitudes of its independent normal modes of vibration. However, the values we are allowed to specify are restricted by the constraint (15.36), which now tells us that all of the Virasoro generators must vanish: $L_n = \widetilde{L}_n = 0$ for every n. Of particular importance are the constraints $L_0 = 0$ and $\widetilde{L}_0 = 0$. These two generators are given by

$$L_0 = -\tfrac{1}{2}\sum_{n=-\infty}^{\infty} \alpha_{-n\,\mu}\alpha^{\mu}_{n} \qquad \widetilde{L}_0 = -\tfrac{1}{2}\sum_{n=-\infty}^{\infty} \widetilde{\alpha}_{-n\,\mu}\widetilde{\alpha}^{\mu}_{n}. \tag{15.78}$$

On separating out the $n=0$ terms which, on account of (15.69) are related to the spacetime momentum, we find

$$M^2 \equiv p_\mu p^\mu = -\frac{4}{\alpha'}\sum_{n=1}^{\infty}\alpha_{-n\,\mu}\alpha_n^\mu = -\frac{4}{\alpha'}\sum_{n=1}^{\infty}\widetilde{\alpha}_{-n\,\mu}\widetilde{\alpha}_n^\mu . \tag{15.79}$$

According to (7.1), this is the equation which gives us the mass M of the string in terms of its vibrational energy. More accurately, we have a pair of equations which, if rewritten as

$$M^2 = -\frac{2}{\alpha'}\sum_{n=1}^{\infty}\left(\alpha_{-n\,\mu}\alpha_n^\mu + \widetilde{\alpha}_{-n\,\mu}\widetilde{\alpha}_n^\mu\right) \tag{15.80}$$

$$\sum_{n=1}^{\infty}\alpha_{-n\,\mu}\alpha_n^\mu = \sum_{n=1}^{\infty}\widetilde{\alpha}_{-n\,\mu}\widetilde{\alpha}_n^\mu \tag{15.81}$$

tell us that the energies of ‘left-moving’ and ‘right-moving’ vibrations must contribute equally to the overall mass.

Much of this analysis also applies to an open string, but there are some significant differences. After gauge fixing, the boundary condition (15.18) becomes $\partial_\sigma X^\mu(\tau,\sigma) = 0$ at $\sigma = 0$ and $\sigma = \ell$. To deal with this, it is convenient to choose $\ell = \pi$, so the range of σ is now $0 \le \sigma \le \pi$. The general solution to the wave equation that satisfies the boundary conditions is

$$\begin{aligned} X^\mu(\tau,\sigma) &= x^\mu + 2\alpha' p^\mu \tau + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2}\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{1}{n}\left[\alpha_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} + \alpha_n^\mu \mathrm{e}^{-\mathrm{i}n(\tau+\sigma)}\right] \\ &= x^\mu + 2\alpha' p^\mu \tau + \mathrm{i}(2\alpha')^{1/2}\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{1}{n}\alpha_n^\mu \mathrm{e}^{-\mathrm{i}n\tau}\cos(n\sigma). \end{aligned} \tag{15.82}$$

Compared with (15.60), there is only one set of coefficients α_n^μ and the factor multiplying p^μ is doubled, so that p^μ is still equal to the spacetime momentum

$$P^\mu = \frac{1}{2\pi\alpha'}\int_0^\pi \mathrm{d}\sigma\,\partial_\tau X^\mu(\tau,\sigma) = p^\mu . \tag{15.83}$$

The expansion (15.82) is a ‘half-range’ Fourier series, which means that we cannot immediately apply the orthogonality relation (15.65) to extract the coefficients α_n^μ. A standard method of dealing with this is to define functions $\mathcal{X}^\mu(\tau,\sigma)$, whose argument σ takes values between 0 and 2π, by

$$\mathcal{X}^\mu(\tau,\sigma) = \begin{cases} X^\mu(\tau,\sigma) & \text{for } 0 \le \sigma \le \pi \\ X^\mu(\tau, 2\pi - \sigma) & \text{for } \pi \le \sigma \le 2\pi \end{cases} \tag{15.84}$$

and their conjugate momenta $\Xi^\mu(\tau,\sigma) = (2\pi\alpha')^{-1}\partial_\tau \mathcal{X}^\mu(\tau,\sigma)$. We get the right answer for α_n^μ by using $\mathcal{X}^\mu$ and Ξ^μ in (15.66) and readers may like to check that (15.67) gives the same result. It is easy to see that $\mathcal{X}^\mu$ and Ξ^ν have the same Poisson bracket as X^μ and Π^ν, so we find again that

$$\{\alpha_m^\mu, \alpha_n^\nu\}_{\mathrm{P}} = \mathrm{i}m\eta^{\mu\nu}\delta_{m,-n}. \tag{15.85}$$

Similarly, we can define extended versions of the components of the energy–momentum tensor

$$\mathcal{T}(w) = -\frac{1}{\alpha'}\,\partial\mathcal{X}_\mu\partial\mathcal{X}^\mu \qquad \widetilde{\mathcal{T}}(\bar{w}) = -\frac{1}{\alpha'}\,\bar{\partial}\mathcal{X}_\mu\bar{\partial}\mathcal{X}^\mu \tag{15.86}$$

which are equal to $T(w)$ and $\widetilde{T}(\bar{w})$ when $\sigma = \operatorname{Re} w$ lies between 0 and π. We find that

$$\mathcal{T}(w) = \sum_{n=-\infty}^{\infty} L_n \mathrm{e}^{\mathrm{i}nw} \qquad \widetilde{\mathcal{T}}(\bar{w}) = \sum_{n=-\infty}^{\infty} L_n \mathrm{e}^{-\mathrm{i}n\bar{w}} \tag{15.87}$$

with a single set of Virasoro generators given by

$$L_n = -\tfrac{1}{2}\sum_{m=-\infty}^{\infty} \alpha_{m\,\mu}\alpha_{n-m}^\mu \tag{15.88}$$

whose Poisson brackets are

$$\{L_m, L_n\}_{\mathrm{P}} = -\mathrm{i}(m-n)L_{m+n}. \tag{15.89}$$

Finally, because of the extra factor of 2 multiplying p^μ in (15.82), we must identify

$$\alpha_0^\mu = (2\alpha')^{1/2}p^\mu \tag{15.90}$$

and the constraint $L_0 = 0$ now gives the mass of an open string as

$$M^2 = -\frac{1}{\alpha'}\sum_{n=1}^{\infty}\alpha_{-n\,\mu}\alpha_n^\mu \tag{15.91}$$

in place of (15.79).

15.2.5 A useful transformation

For some purposes, it is helpful to rewrite our theory in terms of a complex coordinate z, related to w by the conformal transformation

$$z = \mathrm{e}^{-\mathrm{i}w} = \mathrm{e}^{\sigma^2}\mathrm{e}^{-\mathrm{i}\sigma^1}. \tag{15.92}$$

Since this is to be regarded as a conformal transformation, rather than a mere change of coordinates, it involves a change in the worldsheet metric. Let us see

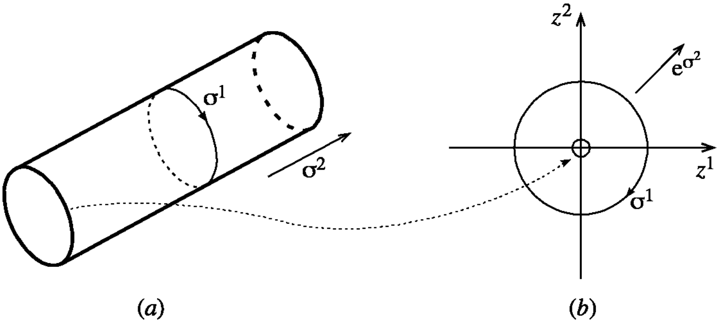


Figure 15.2. The internal geometry of a closed-string worldsheet with two different choices of the metric, related by a conformal transformation. In (*a*), the line element is given by (15.93), while in (*b*) it is given by (15.94). The worldsheet is flat in both cases. In (*b*), it occupies the whole of the complex z plane, except for an infinitesimal disc at the origin, which corresponds to one end of the cylinder in (*a*).

what this means for a closed string. Using the coordinate w, an element of proper distance on the worldsheet is

$$ds^2 = dw\, d\bar{w} = (d\sigma^1)^2 + (d\sigma^2)^2. \tag{15.93}$$

This is a flat, Euclidean metric, so the *internal* geometry of the worldsheet is accurately represented by the long straight cylinder shown in figure 15.2(*a*). (However, according to our discussion at the end of §15.2.2, this does *not* mean that the worldsheet looks like a straight cylinder when embedded in spacetime.) After the conformal transformation, the element of proper distance is

$$ds^2 = dz\, d\bar{z} = (dz^1)^2 + (dz^2)^2 = d\rho^2 + \rho^2 d\theta^2 \tag{15.94}$$

where z^1 and z^2 are the real and imaginary parts of z, while $\rho = e^{\sigma^2}$ is its magnitude and $\theta = -\sigma^1$ is its phase. This is also a flat, Euclidean metric, but now the Euclidean 'time' σ^2 runs in the radial direction, while σ^1 is minus the polar angle. The circular end of the worldsheet at $\sigma^2 \to -\infty$ is an infinitesimal circle at the origin $z = 0$ and the other end, at $\sigma^2 \to +\infty$ is the very large circle at $|z| \to \infty$ (figure 15.2(*b*)).

Written in terms of z, the mode expansion (15.70) is

$$X^\mu(z,\bar{z}) = x^\mu - \tfrac{1}{2} i\alpha' p^\mu (\ln z + \ln \bar{z}) + i\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{n}\left[\alpha_n^\mu z^{-n} + \tilde{\alpha}_n^\mu \bar{z}^{-n}\right]. \tag{15.95}$$

The derivative $\partial_z X^\mu$ is a Laurent series of positive and negative powers of z and $\partial_{\bar{z}} X^\mu$ is a Laurent series in $\bar{z}$. The components of the energy–momentum tensor are also Laurent series, namely

$$T_{zz}=\left(\frac{\mathrm{d}w}{\mathrm{d}z}\right)^2 T_{ww}=-\sum_{n=-\infty}^{\infty}L_n z^{-(n+2)} \tag{15.96}$$

$$T_{\bar{z}\bar{z}}=\left(\frac{\mathrm{d}\bar{w}}{\mathrm{d}\bar{z}}\right)^2 T_{\bar{w}\bar{w}}=-\sum_{n=-\infty}^{\infty}\widetilde{L}_n \bar{z}^{-(n+2)}. \tag{15.97}$$

As an application of this transformation, readers should be able to verify that the quantity $\boldsymbol{T}(\epsilon,\bar{\epsilon})$ defined in (15.59), which generates an infinitesimal conformal transformation is

$$\boldsymbol{T}(\eta,\bar{\eta})=-\oint\frac{\mathrm{d}z}{2\pi}\,\eta(z)T_{zz}(z)-\oint\frac{\mathrm{d}\bar{z}}{2\pi}\,\bar{\eta}(\bar{z})T_{\bar{z}\bar{z}}(\bar{z}) \tag{15.98}$$

where the first integral is over a closed, anticlockwise contour encircling the origin in the z plane and the second is over a closed, anticlockwise contour in the $\bar{z}$ plane. The function $\eta(z)=-\mathrm{i}z\epsilon(w)$ is the small change in z when w changes by a small amount $\epsilon(w)$. On account of Cauchy's theorem, the value of $\boldsymbol{T}(\eta,\bar{\eta})$ is independent of the contour that we use to calculate it, so long as this contour winds once round the origin, where T_{zz} and $T_{\bar{z}\bar{z}}$ have poles. In particular, the value of the original expression (15.59) is independent of τ, which appears on the right-hand side. Note though, that this is true only when X^μ is a solution of the equation of motion (15.35), because this is what makes T holomorphic and $\widetilde{T}$ antiholomorphic. Another, related application is suggested in exercise 15.4.

The same transformation can be used in the case of an open string, whose worldsheet as viewed in the w frame of reference is the flat strip shown in figure 15.3(a). The conformal transformation (15.92) maps it into the lower half of the complex z plane, as shown in figure 15.3(b). The end at $\sigma^2\to-\infty$ is mapped into an infinitesimal semicircle at the origin, while its long edges become the two halves of the real z axis. On the other hand, the extended versions $\mathcal{X}^\mu$ of the fields defined in (15.84) live on the whole complex z plane. Integrating a function of X^μ over σ from 0 to π is equivalent to integrating the corresponding function of $\mathcal{X}^\mu$ around the whole closed contour shown in figure 15.3(b), so Cauchy's theorem can be applied here too.

15.3 Quantization of the Free Bosonic String

The first step in quantizing our classical string is simple enough: in accordance with the principles established in chapters 5 and 7, we promote the coefficients α_n^μ in the mode expansion (15.82) for the open string, or α_n^μ and $\widetilde{\alpha}_n^\mu$ in (15.60) for the closed string, to operators and the Poisson bracket relations (15.68) to the commutation relations

$$[\alpha_m^\mu,\alpha_n^\nu]=[\widetilde{\alpha}_m^\mu,\widetilde{\alpha}_n^\nu]=-m\eta^{\mu\nu}\delta_{m,-n}\qquad[\alpha_m^\mu,\widetilde{\alpha}_n^\nu]=0. \tag{15.99}$$

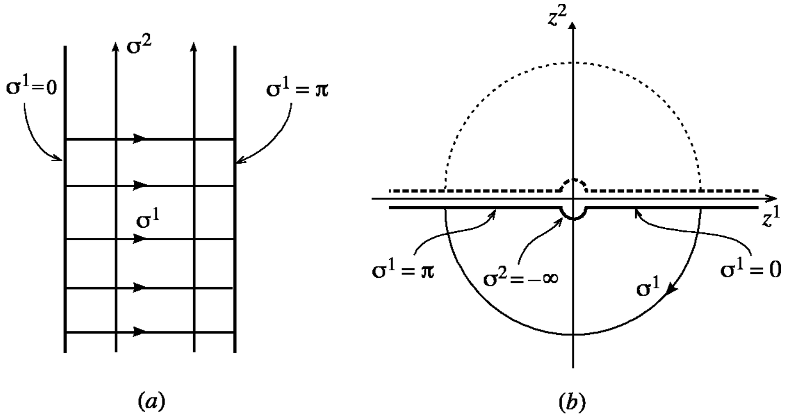


Figure 15.3. The internal geometry of an open-string worldsheet with the same two choices of metric as those used for a closed string in figure 15.2. In (*b*), the worldsheet occupies the lower half of the complex z plane, the short edge at $\sigma^2 \to -\infty$ being an infinitesimal semicircle below the origin. The extended fields defined in (15.84) inhabit the whole z plane, values of σ^1 between π and 2π lying in the upper half plane. Cauchy's theorem can be applied, for example, to the closed contour consisting of the solid semicircle in the lower half plane and the dotted semicircle in the upper half plane.

Superficially, the implications are straightforward. These commutation relations are quite analogous to those of the creation and annihilation operators we have met before. We can expect to be able to interpret these operators as creating and annihilating quanta of energy in the various modes of vibration of the string and, by finding the eigenstates and eigenvalues of the mass^2 operators (15.79) and (15.91), to find out how these quanta of vibrational energy contribute to the overall mass of the string. For a free string, this is all the meaningful information we can ask for, unless we introduce some extra 'internal' degrees of freedom.

There are, however, two outstanding issues, which will require a fair amount of effort to sort out. One is to find a suitable method of dealing with the constraint (15.36), which we now express as $L_n = \widetilde{L}_n = 0$. The other is to determine whether the gauge fixing of §15.2.2 (without which we would not have progressed very far) can be implemented in the quantum theory. This gauge fixing depends on the invariance of the theory under both diffeomorphisms and Weyl transformations, so we must check whether these are still valid symmetries of the quantum theory. What we shall actually do is to check on the validity of the more restricted symmetry of conformal invariance on a flat worldsheet which, as we saw in §15.2.3 is a special combination of a diffeomorphism and a Weyl transformation. The quantum-mechanical properties of the Virasoro generators L_n and $\widetilde{L}_n$ are clearly central to both these issues, so it to these that we turn first.

15.3.1 The quantum Virasoro algebra

Since all of the α_n^μ commute with all of the $\widetilde{\alpha}_n^\mu$, we need consider only one of these sets of operators. We need to decide which of the α_n^μ are to count as creation operators and which as annihilation operators. To this end, let us find the Hamiltonian H, which is the generator of τ translations on the worldsheet. In the case of a closed string, it is given by

$$\begin{aligned} H &= -\int_0^{2\pi} \mathrm{d}\sigma\, \Pi_\mu(\tau,\sigma)\partial_\tau X^\mu(\tau,\sigma) - L \\ &= -\int_0^{2\pi} \mathrm{d}\sigma\, \Pi_\mu(\tau,\sigma)\partial_\tau X^\mu(\tau,\sigma) + \frac{1}{4\pi\alpha'}\int_0^{2\pi} \mathrm{d}\sigma\, \partial_a X_\mu \partial^a X^\mu \\ &= \int_0^{2\pi} \frac{\mathrm{d}\sigma}{2\pi}\left[T(\sigma-\tau) + \widetilde{T}(\sigma+\tau)\right] \\ &= L_0 + \widetilde{L}_0. \end{aligned} \tag{15.100}$$

In the first line, the $-$ sign in the first term again ensures that the *spatial* components $\sum_i \Pi^i \partial_\tau X^i$ appear with the same sign as in the non-relativistic definition (3.14) and the Lagrangian L is the action (15.31) without the τ integral. The commutation relations

$$\left[\alpha_n^\mu, H\right] = n\alpha_n^\mu \tag{15.101}$$

follow from expressing L_0 and $\widetilde{L}_0$ in terms of the α_n^μ and the $\widetilde{\alpha}_n^\mu$ as in (15.78) and from the commutators (15.99). By comparing the signs with the corresponding commutators for the harmonic oscillator (5.60) and (5.61), we see that the α_n^μ are annihilation operators for $n > 0$ and creation operators for $n < 0$, although they are differently normalized from a and $a^\dagger$. (This is evidently consistent with our earlier conclusion that $\alpha_n^{\mu\dagger} = \alpha_{-n}^\mu$ so as to make X^μ real.)

The correspondence (5.37) suggests that the Virasoro generators, which classically have the Poisson-bracket relations (15.77), might in the quantum theory satisfy the commutation relations

$$[L_m, L_n] = (m-n)L_{m+n}. \tag{15.102}$$

I emphasized in chapter 3, however, that although this correspondence is often true, it may not always be. In the case at hand, the generators L_n can be expressed in terms of the α_n^μ, whose commutation relations (15.99) are the basic postulate of our quantization procedure. We can check, therefore, whether the commutation relations (15.102) hold or not. This will prove to be no trivial undertaking, but it is crucial. If these commutation relations turn out not to hold (as they will), then the conformal invariance of our quantum theory is modified in a way which invalidates the gauge fixing that we have taken over from the classical theory. How this comes about, I shall explain in more detail when we have the necessary results in hand.

So much depends on our getting the right answer for $[L_m, L_n]$ that I am going to describe the process in some detail. Readers who are prepared to take my word for the validity of the end result may, however, prefer to skip to equation (15.116), where it is displayed. Consider first the expression for the L_n given in (15.73). In the quantum theory it is ambiguous, because the α_n^μ do not commute, so the meaning of L_n depends on the order of the two αs. As a matter of fact, the basic commutator (15.99) tells us that $[\alpha_m^\mu, \alpha_{n-m}^\nu] = 0$, except when $m = m-n$, or when $n = 0$, so the ambiguity affects only L_0. Moreover, the effect of changing the order of $\alpha_{m\,\mu}$ and α_{-m}^μ is just to add a constant to L_0. Therefore, we can express all the ambiguity by writing

$$L_0 = -\frac{\alpha'}{4} p_\mu p^\mu - \sum_{m=1}^{\infty} \alpha_{-m\,\mu}\alpha_m^\mu + a \tag{15.103}$$

where a is an unknown constant. The first term here is $-\frac{1}{2}\alpha_{0\,\mu}\alpha_0^\mu$, re-expressed via (15.69). The second term is normal-ordered, in the sense we met in §7.2, by making all the annihilation operators stand to the right of the creation operators. It should be clear that the sum of terms in (15.73) with $m \leq -1$ make exactly the same contribution to this normal-ordered expression as the terms with $m \geq 1$.

Next, we can use (15.99) to find, after a little algebra, the commutator

$$\left[L_m, \alpha_n^\mu\right] = -n\alpha_{m+n}^\mu \tag{15.104}$$

which is well defined. This in turn can be used to calculate the commutator $[L_m, L_n]$. The result is

$$[L_m, L_n] = -\tfrac{1}{2}(m-n) \sum_{r=-\infty}^{\infty} \alpha_{m+n-r}^\mu \alpha_{r\,\mu}. \tag{15.105}$$

Except when $n = -m$, this commutator is also well defined and equal to $(m-n)L_{m+n}$, in agreement with (15.102). When $n = -m$, we have

$$\left[L_m, L_{-m}\right] = -m \sum_{r=-\infty}^{\infty} \alpha_{-r}^\mu \alpha_{r\,\mu} \tag{15.106}$$

which is troublesome. To see why, let us express the right-hand side in normal-ordered form, as we did with L_0. We have

$$\begin{aligned}\left[L_m, L_{-m}\right] &= -m\alpha_0^\mu\alpha_{0\,\mu} - m\sum_{r=1}^{\infty}\alpha_{-r}^\mu\alpha_{r\,\mu} - m\sum_{r=1}^{\infty}\alpha_r^\mu\alpha_{-r\,\mu} \\ &= -m\frac{\alpha'}{2}p_\mu p^\mu - 2m\sum_{r=1}^{\infty}\alpha_{-r}^\mu\alpha_{r\,\mu} + md\sum_{r=1}^{\infty} r \\ &= 2m(L_0 - a) + md\sum_{r=1}^{\infty} r \end{aligned} \tag{15.107}$$

where $d = \delta^{\mu}_{\mu}$ is the number of spacetime dimensions. In the second line, I have used the commutator (15.99) to rewrite $\alpha^{\mu}_{r}\alpha_{-r\,\mu}$ as $\alpha^{\mu}_{-r}\alpha_{r\,\mu} - r\delta^{\mu}_{\mu}$. The result is meaningless, because its last term is infinite.

We can obtain a more meaningful answer by taking careful account of the Hilbert space in which the operators are to act. A suitable Hilbert space consists of vectors of the form $|\mathcal{O}; k\rangle$, where k is the spacetime momentum of the string (and so $p^{\mu}|\mathcal{O}; k\rangle = k^{\mu}|\mathcal{O}; k\rangle$), while $\mathcal{O}$ represents the state of the infinite number of oscillators which are its normal modes of vibration. The state $|0; k\rangle$, in which all these oscillators are in their ground states, is annihilated by all the annihilation operators: that is

$$\alpha^{\mu}_{n}|0; k\rangle = 0 \qquad \text{for } n \geq 1. \tag{15.108}$$

A complete basis for the Hilbert space consists of the vectors that we get by acting on $|0; k\rangle$ with any combination of creation operators, which add quanta of vibrational energy. Because our theory is Lorentz invariant, it will be enough to consider the string in the rest frame of its centre of mass or, in other words, to consider just states of the form $|\mathcal{O}; 0\rangle$. Our problem is to find a meaningful interpretation for the quantity

$$\mathcal{L}_m = \left[L_m, L_{-m}\right] - 2m(L_0 - a). \tag{15.109}$$

A few lines of algebra using (15.104) should enable readers to verify that $\mathcal{L}_m$ commutes with all the α^{μ}_{n}. So the action of $\mathcal{L}_m$ on any of our basis vectors, say $\alpha^{\mu_1}_{n_1}\cdots\alpha^{\mu_N}_{n_N}|0; 0\rangle$, where the n_i are all negative, is given by

$$\mathcal{L}_m\alpha^{\mu_1}_{n_1}\cdots\alpha^{\mu_N}_{n_N}|0; 0\rangle = \alpha^{\mu_1}_{n_1}\cdots\alpha^{\mu_N}_{n_N}\mathcal{L}_m|0; 0\rangle \tag{15.110}$$

and we need only to find the value of $\mathcal{L}_m|0; 0\rangle$. This we can do by first finding the actions of the L_m, which are

$$L_0|0; 0\rangle = a|0; 0\rangle \tag{15.111}$$

$$L_{-1}|0; 0\rangle = 0 \tag{15.112}$$

$$L_m|0; 0\rangle = 0 \qquad \text{for } m \geq 1 \tag{15.113}$$

$$L_{-m}|0; 0\rangle = -\tfrac{1}{2}\sum_{r=1}^{m-1}\alpha^{\mu}_{r-m}\alpha_{-r\,\mu}|0; 0\rangle \qquad \text{for } m \geq 2. \tag{15.114}$$

In the case of L_0, I have used the normal-ordered expression (15.103); for all the other L_m, we simply discard all the terms containing either p^{μ} (or α^{μ}_{0}) or an annihilation operator.

For $m \geq 1$, the results (15.111) and (15.113) tell us that $\mathcal{L}_m|0; 0\rangle = L_m L_{-m}|0; 0\rangle$, so we can calculate

$$\mathcal{L}_m|0; 0\rangle = -\tfrac{1}{2}L_m\sum_{r=1}^{m-1}\alpha^{\mu}_{r-m}\alpha_{-r\,\mu}|0; 0\rangle$$

$$
\begin{aligned}
&= -\tfrac{1}{2}\sum_{r=1}^{m-1}\left[\alpha^{\mu}_{r-m}\alpha_{-r\,\mu}L_m + r\alpha^{\mu}_{r-m}\alpha_{m-r\,\mu}\right.\\
&\qquad\qquad\qquad\qquad \left.+(m-r)\alpha^{\mu}_{r}\alpha_{-r\,\mu}\right]|0;0\rangle\\
&= -\tfrac{1}{2}\sum_{r=1}^{m-1}(m-r)\alpha^{\mu}_{r}\alpha_{-r\,\mu}|0;0\rangle\\
&= -\tfrac{1}{2}\sum_{r=1}^{m-1}(m-r)\left[\alpha_{-r\,\mu}\alpha^{\mu}_{r} - r\delta^{\mu}_{\mu}\right]|0;0\rangle\\
&= \tfrac{1}{2}d\sum_{r=1}^{m-1}(m-r)r|0;0\rangle\\
&= \frac{d}{12}m(m^2-1)|0;0\rangle.
\end{aligned}
\tag{15.115}
$$

The first line of this calculation uses the result (15.114); the second uses the commutator (15.104) twice; the third discards terms in which annihilation operators act on $|0;0\rangle$; the fourth uses the commutator (15.99) and the fifth discards further terms in which annihilation operators act on $|0;0\rangle$.

For $m \le -1$, a similar calculation gives the same result and we also get the same result when $m = 0$, because $[L_0, L_0] = 0$. By virtue of (15.110), we see that $\mathcal{L}_m|\Psi\rangle = (d/12)m(m^2-1)|\Psi\rangle$ when $|\Psi\rangle$ is *any* vector in the Hilbert space we specified, so we can simply take $\mathcal{L}_m = (d/12)m(m^2-1)$. This gives us the value of $\left[L_m, L_{-m}\right]$ and we discovered earlier that $[L_m, L_n] = (m-n)L_{m+n}$ when n is not equal to $-m$. Thus, we can finally write the commutation relations of the quantum Virasoro algebra as

$$
[L_m, L_n] = \left[\frac{m(m^2-1)}{12}c - 2ma\right]\delta_{m,-n} + (m-n)L_{m+n} \tag{15.116}
$$

where the constant c is known, in a general conformal field theory, as the *central charge*. For this particular theory, of course, we have found that c is equal to the number of spacetime dimensions d.

Let us take stock of what we have learned. The commutation relations of the quantum Virasoro algebra differ from the Poisson-bracket relations of the classical theory by the first term in (15.116), which involves two constants c and a. The constant a arises from the normal ordering of L_0. We have met a similar constant before, in (7.21), where it represented the energy of the vacuum state. The value of this vacuum energy is essentially a matter of convention, so we were entitled to set it to zero by discarding the constant. Here, the ‘vacuum energy’ of the worldsheet field theory is the mass of the string in its vibrational ground state, whose value is by no means a matter of convention. The classical formula (15.79) for the mass of a closed string was obtained from the constraints

$L_0 = \widetilde{L}_0 = 0$ and its quantum-mechanical version contains two normal-ordering constants a and $\widetilde{a}$, because exactly the same considerations apply to the $\widetilde{L}_n$ as to the L_n. If we want to find a reliable prediction for the mass of the string (as we assuredly do!), then the correct values of a and $\widetilde{a}$ must be found. For the moment, I shall simplify matters by *assuming* that $a = \widetilde{a} = 0$. Later on, we shall determine the correct quantum-mechanical mass formula from considerations having to do with the internal consistency of our quantization procedure. With $a = 0$, the commutation relations are

$$[L_m, L_n] = \frac{m(m^2-1)}{12}\, c\, \delta_{m,-n} + (m-n)L_{m+n} \tag{15.117}$$

and it is in this form that they are most often quoted.

The modified commutation relations imply a modification in the conformal invariance of the quantum theory, which is most conveniently treated from the point of view of the complex coordinate z introduced in §15.2.5. The conformal transformation of a quantum-mechanical operator A is given by

$$A' = \exp[-\mathrm{i}\boldsymbol{T}(\eta, \bar{\eta})]\, A \exp[\mathrm{i}\boldsymbol{T}(\eta, \bar{\eta})] \tag{15.118}$$

where $\boldsymbol{T}(\eta, \bar{\eta})$ is the quantum version of (15.98), which is analogous, for example, to the spacetime translation of (12.110). The infinitesimal version is

$$\delta A = -\mathrm{i}\,[\boldsymbol{T}(\eta, \bar{\eta}), A]\,. \tag{15.119}$$

To illustrate how this works, consider the operator $\partial X^\mu(z)$, where I now use ∂ to mean $\partial/\partial z$. Because ∂X^μ is independent of $\bar{z}$, its transformation is generated by $\boldsymbol{T}(\eta, 0)$. We use the Laurent expansions

$$\partial X^\mu(z) = -\mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{n=-\infty}^{\infty} \alpha_n^\mu z^{-(n+1)} \qquad T_{zz}(z) = -\sum_{m=-\infty}^{\infty} L_m z^{-(m+2)}$$

$$\eta(z) = \sum_{\ell=-\infty}^{\infty} \eta_\ell z^{-\ell}$$

and Cauchy's theorem in the form $\oint \mathrm{d}z\, z^{-n} = 2\pi\mathrm{i}\,\delta_{n,1}$ to calculate

$$\begin{aligned}
\delta\left(\partial X^\mu(z)\right) &= -\mathrm{i}\left[\boldsymbol{T}(\eta, 0), \partial X^\mu(z)\right] \\
&= -\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\ell,m,n} \oint \frac{\mathrm{d}z'}{2\pi}\, z'^{-(\ell+m+2)} z^{-(n+1)} \eta_\ell \left[L_m, \alpha_n^\mu\right] \\
&= \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\ell,m,n} \eta_\ell\, \alpha_{m+n}^\mu n\, \delta_{\ell+m+1,0} z^{-(n+1)} \\
&= \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\ell,n} \eta_\ell\, \alpha_n^\mu (\ell+n+1) z^{-(\ell+n+2)} \\
&= \partial\left[\eta(z)\partial X^\mu(z)\right].
\end{aligned} \tag{15.120}$$

The result is what we might expect from our classical discussion of conformal invariance in §15.2.3. That is to say

$$\begin{aligned}\delta\left(\partial X^\mu(z)\right) &= \partial X^\mu\left(z+\eta(z)\right) - \partial X^\mu(z) = \partial\left[X^\mu\left(z+\eta(z)\right) - X^\mu(z)\right]\\ &\simeq \partial\left[\eta(z)\partial X^\mu(z)\right].\end{aligned} \qquad (15.121)$$

For the energy–momentum tensor $T_{zz}(z) = -(1/\alpha')\partial X_\mu(z)\partial X^\mu(z)$, the classical theory would lead us to expect

$$\delta T_{zz}(z) = -\frac{2}{\alpha'}\,\partial X_\mu(z)\delta\left(\partial X^\mu(z)\right) = 2\left(\partial\eta(z)\right)T_{zz}(z) + \eta(z)\partial T_{zz}(z). \qquad (15.122)$$

However, a calculation similar to (15.120) gives the result

$$\delta T_{zz}(z) = -\frac{c}{12}\partial^3\eta(z) + 2\left(\partial\eta(z)\right)T_{zz}(z) + \eta(z)\partial T_{zz}(z) \qquad (15.123)$$

so $T_{zz}(z)$ does not change as it would under a classical conformal transformation.

Whether we regard this as a breakdown of conformal invariance in the quantum theory is a matter of taste. The transformations generated by the quantum commutation relations (15.104) and (15.117) do constitute a symmetry of the quantum theory, and there is a set of infinitely many conserved quantities associated with this symmetry (see exercise 15.5). The usual practice is to call this symmetry conformal invariance, but it is not quite the same as the classical invariance. The extra terms in (15.117) and (15.123), proportional to the central charge, are said to arise from an *anomaly*, of the kind that I mentioned briefly in chapter 9. The one we have discovered here is the *conformal* or *Virasoro anomaly*. In the next subsection, we shall see that it is sufficient to invalidate gauge fixing in the quantum theory, except under special (and somewhat curious) circumstances. Incidentally, the scale invariance that we studied in §11.6 in connection with critical phenomena is a special case of conformal invariance, corresponding to a rescaling of z by a constant factor, $f(z) = \ell z$. The fact that the critical exponents of the 'classical' Ginzburg–Landau theory are modified when statistical fluctuations are taken into account by what amounts to a quantum field theory can be understood as a manifestation of this same conformal anomaly.

15.3.2 Quantum gauge fixing

At the beginning of this section, I pointed out that there are two issues we must confront in developing a reliable quantization procedure. One is how to deal with the constraint $T_{ab} = 0$ which, in terms of the mode expansion becomes $L_n = 0$; the other is to determine how, if at all, we can fix the gauge so as to be able to work with a flat worldsheet. Several different approaches to string quantization have been investigated, within which these issues can be addressed. The most reliable, and the one I am going to describe, is known as BRST quantization, after

C Becchi, A Rouet, R Stora and I V Tyutin who developed a similar strategy for the quantization of non-Abelian gauge theories. It relies on the possibility of expressing matrix elements in terms of path integrals, analogous to the particle propagator (15.7). Consider, in particular the object

$$Z = \int \mathcal{D}X\mathcal{D}\gamma \exp[\mathrm{i}S(X, \gamma)] \tag{15.124}$$

where $S(X, \gamma)$ is the original action (15.15). This does not correspond directly to any observable quantity (we are not going to impose the constraint that the worldsheet begin and end at some specified locations in spacetime), though is somewhat analogous to a partition function in statistical mechanics. Its relevance to our present purpose is that we can carry out various manipulations with this path integral, which can subsequently be reinterpreted in terms of operators acting in the Hilbert space of string states.

In the path-integral context, we can attempt to fix the gauge by means of the Fadeev–Popov method to which I alluded in §9.5. There, the details were not greatly relevant to our discussion (though they are important to anyone engaged in 'real life' calculations). Here, they play a central role, so I will treat them much more explicitly. The essential point is that the integral over γ_{ab} includes integrals over many gauge degrees of freedom (the diffeomorphisms and Weyl transformations) which do not change the value of S. They ought, therefore, to contribute only a constant factor to Z, say $\mathcal{V}_{\text{gauge}}$, which represents the 'volume' of the space over which the gauge variables are integrated. In fact, the result we shall obtain has the general form

$$Z = \mathcal{V}_{\text{gauge}} \int \mathcal{D}X\mathcal{D}b\mathcal{D}c \exp\left[\mathrm{i}S(X, \eta) + \mathrm{i}S_{\mathrm{g}}(b, c)\right] \tag{15.125}$$

where b and c are 'ghost' fields, whose presence I shall explain later on. Of course, $S(X, \eta)$ is the gauge-fixed action with the flat worldsheet metric η_{ab}. Before embarking on the derivation of this result, let us consider the status of the constraint. From the previous subsection, it is clear that we cannot set $L_n = 0$ in the quantum theory. If we did, the commutation relation (15.104) would require us also to set all the α_n^μ equal to zero, so vibrations of the string would be entirely forbidden. Apart from L_0 which, as we have seen, needs special treatment, each term in the sums (15.73) is a product of two creation or annihilation operators which affect the states of two different oscillators. It follows that for any vector $|\Psi\rangle$ in the Hilbert space, the vector $L_n|\Psi\rangle$ either vanishes or represents a state in which two of the oscillators have different energies from those that they have in the state $|\Psi\rangle$. Thus, $L_n|\Psi\rangle$ is orthogonal to $|\Psi\rangle$ and the expectation value $\langle\Psi|L_n|\Psi\rangle$ is zero. This is the way in which the constraint is realized in the theory as we have studied it so far. From the point of view of the path integral (15.124), the integral over γ_{ab} provides, roughly speaking, a functional δ-function, which enforces the constraint. In the gauge-fixed form (15.125), this δ-function has disappeared, but provided that the new integral really is equal to the old one, its effect is taken care of in the new theory of the X^μ and the ghosts. The upshot is that in this modified theory, we do not have to take explicit account of the constraints. We shall find, however, that they reappear in the guise of what is called the *BRST cohomology*.

Here is the derivation of (15.125). Our quantum theory is supposedly invariant under a gauge transformation specified by three functions $s^a(\tau,\sigma)$ and $\omega(\tau,\sigma)$, corresponding to a coordinate transformation $\tau' = s^0(\tau,\sigma)$, $\sigma' = s^1(\tau,\sigma)$ and a Weyl transformation (15.26). I denote these three functions collectively by g (for 'gauge transformation'). The effect of a gauge transformation on the fields X and γ is to change is to change them into a new set of functions, which I denote by X^g and γ^g; for example,

$$(\gamma^g)^{a'b'}(\tau',\sigma') = \mathrm{e}^{-\omega}\frac{\partial s^{a'}}{\partial\sigma^a}\frac{\partial s^{b'}}{\partial\sigma^b}\gamma^{ab}(\tau,\sigma). \tag{15.126}$$

By making a suitable gauge transformation, any metric can be reduced to the flat metric η^{ab}. We can turn this round and say that the metric γ inside the original path integral (15.124) is obtained from η by a gauge transformation g, so $\gamma = \eta^g$. Similarly, the fields X are obtained from some other fields X' by means of the same gauge transformation, so $X = X'^g$. The gauge volume $\mathcal{V}_{\text{gauge}}$ that we would like to extract as an overall factor is $\mathcal{V}_{\text{gauge}} = \int \mathcal{D}g\, 1$, and we might think of doing this by exchanging the integral over γ for an integral over g. That is to say, the integral can be written

$$Z = \int \mathcal{D}g \int \mathcal{D}X'^g\ J\ \exp\left[\mathrm{i}S(X'^g,\eta^g)\right] \tag{15.127}$$

where $J = \det(\delta\gamma/\delta g)$ is the Jacobian for this change of variables. If the theory really is gauge invariant, then integral $\int \mathcal{D}X'^g\ J\ \exp(\mathrm{i}S)$ is independent of g, and we have

$$Z = \mathcal{V}_{\text{gauge}} \int \mathcal{D}X\ J\ \exp\left[\mathrm{i}S(X,\eta)\right] \tag{15.128}$$

after dropping the prime from the dummy integration variable X'.

The difficulty here is that the Jacobian J is not easy to determine, so we proceed indirectly, as follows. Define a function $\Delta(\gamma)$ by

$$\Delta(\gamma)^{-1} = \int \mathcal{D}g\, \delta\left(\gamma - \eta^g\right). \tag{15.129}$$

This is an integral over *all* sets of gauge functions g, but the functional δ-function vanishes except when $\eta^g = \gamma$. Clearly, the integral

$$Z = \int \mathcal{D}X\mathcal{D}\gamma\mathcal{D}g\ \Delta(\gamma)\delta\left(\gamma - \eta^g\right)\ \exp\left[\mathrm{i}S(X,\gamma)\right] \tag{15.130}$$

is equal to the original one, because the extra factor in the integrand is equal to 1. We can use the δ-function to carry out the integral over γ and write

$$\begin{aligned} Z &= \int \mathcal{D}g\mathcal{D}X\, \Delta(\eta^g) \exp\left[\mathrm{i}S(X, \eta^g)\right] \\ &= \int \mathcal{D}g \left\{ \int \mathcal{D}X'^g\, \Delta(\eta^g) \exp\left[\mathrm{i}S(X'^g, \eta^g)\right]\right\} \\ &= \mathcal{V}_{\text{gauge}} \int \mathcal{D}X\, \Delta(\eta) \exp\left[\mathrm{i}S(X, \eta)\right] \end{aligned} \tag{15.131}$$

provided, again, that the integral in the curly brackets is genuinely independent of g.

Evidently, the Jacobian we needed to find is $J = \Delta(\eta)$ and we ought to be able to calculate it from the definition (15.129). We have

$$\Delta(\eta)^{-1} = \int \mathcal{D}g\, \delta(\eta - \eta^g) \tag{15.132}$$

and, since the δ-function vanishes except when $\eta^g = \eta$, we need to know η^g only when g is infinitesimal. More specifically, in the gauge transformation (15.126), we take $\omega(\tau, \sigma)$ to be infinitesimal, and the coordinate transformation functions to be $s^a(\tau, \sigma) = \sigma^a + \epsilon^a(\tau, \sigma)$, where the ϵ^a are also infinitesimal. With a little rearrangement, we find that the infinitesimal change in γ^{ab} is

$$\begin{aligned} \delta\gamma^{ab} &= -\omega\gamma^{ab} + \gamma^{ac}\partial_c\epsilon^b + \gamma^{bc}\partial_c\epsilon^a - \left(\partial_c\gamma^{ab}\right)\epsilon^c \\ &= -\omega\gamma^{ab} + \nabla^a\epsilon^b + \nabla^b\epsilon^a \end{aligned} \tag{15.133}$$

where ∇ is the covariant derivative associated with the metric γ. For our immediate purpose, with $\gamma^{ab} = \eta^{ab}$, this implies that

$$\eta^{ab} - (\eta^g)^{ab} = \omega\eta^{ab} - \partial^a\epsilon^b - \partial^b\epsilon^a \tag{15.134}$$

but the more general version will shortly be useful also. The integral over g in (15.132) is now written more explicitly as an integral over the three functions ω and ϵ^a, and the δ-function can be dealt with by using a functional generalization of the integral representation (A.11):

$$\Delta(\eta)^{-1} = \int \mathcal{D}\omega\mathcal{D}\epsilon \int \mathcal{D}\beta\, \exp\left[\frac{\mathrm{i}}{4\pi}\int \mathrm{d}^2\sigma\beta_{ab}\left(\partial^a\epsilon^b + \partial^b\epsilon^a - \omega\eta^{ab}\right)\right]. \tag{15.135}$$

Because the metric is symmetric, there is really one δ-function for each of the independent components η^{00}, η^{11} and $\eta^{10} = \eta^{01}$. Correspondingly, there are three new integration variables, which are the independent components of a *symmetric* tensor field β_{ab}; the factor $1/4\pi$ simply sets a convenient normalization for β_{ab}.

The integral over ω is the integral representation of the δ-function $\delta\left(\beta_{ab}\eta^{ab}\right)$, so we can simplify our result to the form

$$\Delta(\eta)^{-1} = \int \mathcal{D}\epsilon \int \mathcal{D}\beta \, \exp\left[\frac{\mathrm{i}}{2\pi}\int \mathrm{d}^2\sigma \beta_{ab}\partial^a\epsilon^b\right] \tag{15.136}$$

on the understanding that the β integral is now over the two independent components of a tensor field which is both symmetric and *traceless*, in the sense that $\eta^{ab}\beta_{ab} = \beta_{00} - \beta_{11} = 0$. I have also used the fact that $\beta_{ab}\partial^a\epsilon^b = \beta_{ab}\partial^b\epsilon^a$ on account of the symmetry of β_{ab}. Finally, we obtain an expression for $\Delta(\eta)$ itself by making use of the properties of Grassmann variables discussed in appendix A. It is given by

$$\Delta(\eta) = \int \mathcal{D}b\mathcal{D}c \exp\left[\frac{\mathrm{i}}{2\pi}\int \mathrm{d}^2\sigma \, b_{ab}\partial^a c^b\right] \tag{15.137}$$

where b_{ab} and c^a are fields of Grassmann variables, each having two independent components. They can be thought of as the fields associated with a fictitious set of 'particles' living on the worldsheet, which are conventionally described as *Fadeev–Popov ghosts*. The gauge-fixed partition function is therefore indeed given by (15.125), with the ghost action

$$S_{\mathrm{g}}(b, c) = \frac{1}{2\pi}\int \mathrm{d}\tau\mathrm{d}\sigma \, b_{ab}\partial^a c^b. \tag{15.138}$$

The value of the analysis we have been through lies not so much in the partition function itself, but rather in the discovery that we are allowed to fix the gauge in the quantum theory, at the price of dealing with a gauge-fixed theory which includes not only the original fields X^μ but also the ghost fields b and c. This is true, at least, if we can resolve a question that hangs over our derivation. Namely, we must assure ourselves that the quantity in curly brackets in (15.131) is really independent of g, so that $\mathcal{V}_{\text{gauge}}$ can validly be extracted as an overall factor. From time to time, I have given fairly strong hints that this is not in fact so, and we are now in a position to learn the uncomfortable truth of the matter.

15.3.3 The critical spacetime dimension

Now that we know how to express the Jacobian $\Delta(\eta^g)$ in terms of a path integral over ghost fields, the question that confronts us is whether the object

$$Z^g = \int \mathcal{D}X^g\mathcal{D}b^g\mathcal{D}c^g \, \exp\left[\mathrm{i}S(X^g, \eta^g) + \mathrm{i}S_{\mathrm{g}}(b^g, c^g, \eta^g)\right] \tag{15.139}$$

is independent of g. (Note that $S_{\mathrm{g}}(b^g, c^g, \eta^g)$ is not given exactly by the expression (15.138) because η^g is not equal to η; I shall return to this point shortly.) If both coordinate transformations and Weyl transformations are valid symmetries of the quantum theory, as they are of the classical theory, then the gs can be removed from the right-hand side of (15.139) simply by making a gauge transformation. For the classical theory, we saw in (15.22) that $\nabla^a T_{ab} = 0$ as a consequence of diffeomorphism invariance and in (15.28) that $T_a^a = 0$ as a consequence of invariance under Weyl transformations. If both symmetries remain valid in the quantum theory, then these two properties of the energy–momentum tensor should also remain valid, and this is the crucial point that we are going to check. The possibility that one or other of these properties might fail arises from operator-ordering ambiguities such as we met in §15.3.1. Let us assume (as is in fact the case) that the products of operators contained in T_{ab} can be ordered in such a way that $\nabla^a T_{ab} = 0$. We need to know whether T_a^a will then also vanish. On a flat worldsheet it does. In terms of the complex coordinates w and $\bar{w}$, the divergence $\partial^a T_{aw}$, for example, is given by

$$\partial^a T_{aw} = \partial^w T_{ww} + \partial^{\bar{w}} T_{\bar{w}w} = -2\left[\partial_{\bar{w}} T_{ww} + \partial_w T_{\bar{w}w}\right] \tag{15.140}$$

if we take account of the metric (15.44). The operator ordering that makes the Virasoro generators L_n and hence the energy–momentum tensor well defined does not affect the fact that $\partial_{\bar{w}} T_{ww} = \partial_w T_{\bar{w}w} = 0$. It also does not require us to introduce a non-zero value for $T_{\bar{w}w}$, which is proportional to T_a^a, so both properties $\partial^a T_{aw} = 0$ and $T_{\bar{w}w} = 0$ can consistently be maintained on the flat worldsheet. Thus, if T_a^a is non-zero on a curved worldsheet, then it must be proportional to some scalar quantity that vanishes in the limit of a flat worldsheet. The only available scalar field with the right dimensions is the two-dimensional Ricci scalar R, so we must have

$$T_a^a = \lambda R \tag{15.141}$$

where λ is a constant. Our problem reduces, then, to determining the value of λ. Only if $\lambda = 0$ are both diffeomorphism invariance and Weyl invariance valid quantum symmetries.

To calculate λ, let us use coordinates z and $\bar{z}$ such the metric has the form

$$\gamma_{ab} = \exp[\Omega(z,\bar{z})]\eta_{ab} \qquad \eta = -\frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{15.142}$$

On a flat, Euclidean worldsheet these become the complex coordinates introduced in (15.92). Using this metric and the result of exercise 15.1, we find

$$T_a^a = -4\mathrm{e}^{-\Omega} T_{\bar{z}z} \qquad R = 4\mathrm{e}^{-\Omega}\bar{\partial}\partial\Omega \tag{15.143}$$

and so (15.141) becomes

$$T_{\bar{z}z} = -\lambda\bar{\partial}\partial\Omega. \tag{15.144}$$

On the other hand, a few lines of algebra using the result of exercise 15.6 shows that the equation $\nabla^a T_{az} = 0$ becomes

$$\bar{\partial} T_{zz} + \partial T_{\bar{z}z} = (\partial\Omega) T_{\bar{z}z} \tag{15.145}$$

or, on account of (15.144),

$$\bar{\partial} T_{zz} = \lambda \left[\bar{\partial}\partial^2\Omega - (\partial\Omega)(\bar{\partial}\partial\Omega) \right]. \tag{15.146}$$

This equation can be used to relate λ to the central charge of the theory on a flat worldsheet if we consider the change that arises from an infinitesimal Weyl transformation. According to (15.26), the change in Ω is just ω, so the change in T_{zz}, which I denote by $(\delta T_{zz})_{\mathrm{Weyl}}$, satisfies

$$\begin{aligned} \bar{\partial}(\delta T_{zz})_{\mathrm{Weyl}} &= \lambda \left[\bar{\partial}\partial^2\omega - (\partial\omega)(\bar{\partial}\partial\Omega) - (\partial\Omega)(\bar{\partial}\partial\omega) \right] \\ &= \lambda\bar{\partial} \left[\partial^2\omega - (\partial\omega)(\partial\Omega) \right]. \end{aligned} \tag{15.147}$$

In fact, since $(\delta T_{zz})_{\mathrm{Weyl}}$ must vanish when $\omega = 0$, we can write

$$(\delta T_{zz})_{\mathrm{Weyl}} = \lambda \left[\partial^2\omega - (\partial\omega)(\partial\Omega) \right] = \lambda\partial^2\omega \tag{15.148}$$

where the last expression applies to a flat worldsheet with $\Omega = 0$.

For the flat worldsheet, $(\delta T_{zz})_{\mathrm{Weyl}}$ can be found from the conformal transformation (15.123). Remember that a conformal transformation is a special combination of coordinate and Weyl transformations. It changes the worldsheet metric in such a way that the components of the new metric in the new coordinates are the same as those of the old metric in the old coordinates. The effect of an infinitesimal coordinate transformation $z' = z + \eta(z)$ and $\bar{z}' = \bar{z} + \bar{\eta}(\bar{z})$ together with Weyl rescaling by a factor e^{ω} is to replace the line element $\mathrm{d}s^2 = \mathrm{d}z\mathrm{d}\bar{z}$ with

$$\mathrm{d}s'^2 = \mathrm{e}^{\omega}\mathrm{d}z\mathrm{d}\bar{z} = \mathrm{e}^{\omega} \left(\frac{\mathrm{d}z'}{\mathrm{d}z} \frac{\mathrm{d}\bar{z}'}{\mathrm{d}\bar{z}} \right)^{-1} \mathrm{d}z'\mathrm{d}\bar{z}' \simeq (1 + \omega - \partial\eta - \bar{\partial}\bar{\eta})\mathrm{d}z'\mathrm{d}\bar{z}' \tag{15.149}$$

which is equal to $\mathrm{d}z'\mathrm{d}\bar{z}'$ if we take $\omega(z, \bar{z}) = \partial\eta(z) + \bar{\partial}\bar{\eta}(\bar{z})$. (Please do not confuse $\eta(z)$, which is an infinitesimal change of coordinates, with η_{ab} in (15.142), which is the Minkowskian metric.) Now, the classical conformal transformation (15.122) for T_{zz} was obtained for the specific example of the X^{μ} field theory, but it simply reflects the coordinate transformation of a rank $\binom{0}{2}$ tensor, and is also valid for the combined energy–momentum tensor of the X^{μ} and the ghosts b and c, which concerns us here. The anomalous first term of the quantum conformal transformation (15.123) can therefore be identified as arising from a Weyl transformation. Thus we have

$$(\delta T_{zz})_{\mathrm{Weyl}} = -\frac{c}{12}\,\partial^3\eta(z) = -\frac{c}{12}\,\partial^2\omega(z, \bar{z}) = \lambda\partial^2\omega(z, \bar{z}). \tag{15.150}$$

We have discovered that

$$\lambda = -\frac{c}{12} \tag{15.151}$$

where c is the central charge of the combined theory of spacetime coordinates X^{μ} and the Fadeev–Popov ghosts.

To determine the total central charge, we need to know the energy–momentum tensor of the ghosts. This can be found from the definition (15.21) if we use the covariant version of the ghost action (15.138), which contains the non-Minkowskian metric $\gamma = \eta^{g}$. By retaining the general form of the metric variation (15.133), we obtain an action that can be expressed in the form

$$S_{\mathrm{g}}(b, c) = \frac{1}{4\pi}\int \mathrm{d}\tau\mathrm{d}\sigma\,(-\gamma)^{1/2} b_{ab}\left[\nabla^{a}c^{b} + \nabla^{b}c^{a} - \gamma^{ab}\nabla_{c}c^{c}\right]. \qquad (15.152)$$

The factor containing c^{a} is written so as to be symmetric and traceless; it has just two independent components, matching the two degrees of freedom in b_{ab}, which is also symmetric and traceless. Because b_{ab} is traceless, the term $-b_{ab}\gamma^{ab}\nabla_{c}c^{c}$ vanishes, but its variation $-b_{ab}\delta\gamma^{ab}\nabla_{c}c^{c}$ is non-zero. Relative to the $(z, \bar{z})$ coordinates on the flat worldsheet, each of the ghost fields has two independent components, namely

$$b \equiv b_{zz} \qquad \widetilde{b} \equiv b_{\bar{z}\bar{z}} \qquad c \equiv c^{z} \qquad \widetilde{c} \equiv c^{\bar{z}}. \qquad (15.153)$$

With a little patience, readers should find it possible to verify that the Euclidean action for these ghosts is

$$S_{\mathrm{E}}^{(\mathrm{g})} = \frac{1}{2\pi}\int \mathrm{d}z\mathrm{d}\bar{z}\,\left(b\bar{\partial}c + \widetilde{b}\partial\widetilde{c}\right) \qquad (15.154)$$

and that their energy–momentum tensor has the two independent components

$$T_{zz}^{(\mathrm{g})}(z) = 2b(z)\partial c(z) + [\partial b(z)]c(z) \qquad T_{\bar{z}\bar{z}}^{(\mathrm{g})}(\bar{z}) = 2\widetilde{b}(\bar{z})\bar{\partial}\widetilde{c}(\bar{z}) + [\bar{\partial}\widetilde{b}(\bar{z})]\widetilde{c}(\bar{z}) \qquad (15.155)$$

after taking account of the equations of motion $\bar{\partial}b = \bar{\partial}c = \partial\widetilde{b} = \partial\widetilde{c} = 0$ implied by (15.154). The commutation relations of the Virasoro algebra of the ghost theory can be found by the same method that we used to derive (15.116). Rather than embark on another lengthy calculation, I shall simply quote the result that $c^{(\mathrm{g})} = -26$. For the combined theory, we have the Virasoro generators $L_n = L_n^{(X)} + L_n^{(\mathrm{g})}$ and, since $L_m^{(X)}$ commutes with $L_n^{(\mathrm{g})}$ the commutation relations for the combined algebra are

$$\begin{aligned}[L_m, L_n] = &\left[\frac{m(m^2-1)}{12}\left(c^{(X)} + c^{(\mathrm{g})}\right) - 2m\left(a^{(X)} + a^{(\mathrm{g})}\right)\right]\delta_{m,-n} \\ &+ (m-n)L_{m+n}.\end{aligned} \qquad (15.156)$$

The anomalous term vanishes if $a^{(X)} + a^{(\mathrm{g})} = 0$ and $c = c^{(X)} + c^{(\mathrm{g})} = 0$. We can arrange for the normal-ordering constants $a^{(X)}$ and $a^{(\mathrm{g})}$ to add to zero simply by specifying their values as part of our quantization procedure. To be definite, I shall set $a^{(X)} = a^{(\mathrm{g})} = 0$, although the individual values do not really matter. On the other hand, the net central charge $c = d - 26$ is fixed by the commutation relations of creation and annihilation operators.

At this point, two strategies would seem to present themselves. If c does not vanish, we can partially fix the gauge by using diffeomorphism invariance to reduce the worldsheet metric to the 'conformally flat' form (15.142), but we cannot remove the conformal factor e^{Ω} because Weyl invariance is not valid in the quantum theory. This means that the quantum theory contains an extra field $\Omega(z, \bar{z})$, known as the *dilaton* on account of the alternative term 'dilation' for a Weyl transformation. The resulting theory, known as the theory of *non-critical strings*, has been investigated but, to the best of my knowledge, one cannot be sure that it makes good mathematical sense. At any rate, I shall have no more to say about it here. The second strategy, which is more prominent in the string-theory literature, is to suppose that spacetime has as many dimensions as are needed to make the central charge vanish. In the case of the bosonic string that we have studied so far, this *critical dimension* is $d = 26$. One way of accounting for the fact that we perceive only four of these dimensions is to invoke the Kaluza-Klein idea (§8.5) that the extra ones are compactified with a very small size. Another possibility makes use of the idea of 'D-branes', on which I shall touch later.

This value, $d = 26$, for the critical dimension of the bosonic string is our first major result. Clearly, the conclusion that the theory makes sense only if spacetime has 26 dimensions has far-reaching consequences. Readers may well gain the impression that this conclusion rests on a rather inconsequential technicality—the failure of Weyl invariance as a quantum symmetry—which we had to work rather hard to uncover. A lot of difficulty might perhaps be avoided if we were to turn a blind eye to this technical hitch and proceed to develop our theory in 4 spacetime dimensions, in the hope that nothing serious will go wrong in the end. I should emphasize, therefore, that this has been tried and it does not work. Several different approaches to quantizing the string have been developed over many years and all of them produce inconsistencies of one kind or another unless there are exactly 26 of the coordinate fields X^{μ}. In one way or another, proper account must be taken of this fact if further progress is to be possible.

15.3.4 The ghost Hilbert space

We shall naturally need to know something about the quantum-mechanical properties of the ghost fields, in particular the nature of the Hilbert space in which they act. Here, I summarize the essential results for the right-moving fields b and c of a closed string; the left-moving fields $\widetilde{b}$ and $\widetilde{c}$ form an identical structure and the differences for an open string are exactly parallel to those we discussed in the case of the X^{μ} field theory. I shall omit details of most of the derivations, which readers who have progressed this far should find to be matters of (possibly tiresome) routine.

The right-moving part of the action (15.154) is invariant under a conformal transformation analogous to (15.49), where the ghost fields transform as

$$b'(z) = \left(\frac{\mathrm{d}f}{\mathrm{d}z}\right)^{h_b} b\left(f(z)\right) \qquad c'(z) = \left(\frac{\mathrm{d}f}{\mathrm{d}z}\right)^{h_c} c\left(f(z)\right) \tag{15.157}$$

provided that $h_b + h_c = 1$. In conformal field theory, these indices are called the *conformal weights* of the fields. The values of these conformal weights can be found from the requirement that the transformation be generated by the energy–momentum tensor (15.155). One finds $h_b = 2$ and $h_c = -1$, which again reflect the coordinate transformations of a rank $\binom{0}{2}$ and a rank $\binom{1}{0}$ tensor field. We adopt the mode expansions

$$b(\tau,\sigma) = -\sum_{n=-\infty}^{\infty} b_n \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} = -\sum_{n=-\infty}^{\infty} b_n \mathrm{e}^{\mathrm{i}nw} \tag{15.158}$$

$$c(\tau,\sigma) = -\sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} = -\sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{i}nw}. \tag{15.159}$$

Compared with the expansions (15.60) of the X^μ, the missing factor of $1/n$ arises from the fact that the Lagrangian has only one derivative. Since the ghosts are fermions, the expansion coefficients have anticommutation relations, which are

$$\{b_m, c_n\} = \delta_{m,-n} \qquad \{b_m, b_n\} = \{c_m, c_n\} = 0. \tag{15.160}$$

Taking into account the conformal transformations (15.157), the Laurent series for $b(z)$ and $c(z)$ are

$$b(z) = \sum_{n=-\infty}^{\infty} b_n z^{-(n+2)} \qquad c(z) = \sum_{n=-\infty}^{\infty} c_n z^{-(n-1)}. \tag{15.161}$$

For the Virasoro generators we have

$$L_n^{(\mathrm{g})} = \sum_{m=-\infty}^{\infty} (2n-m) : b_m c_{n-m} : -\delta_{n,0}. \tag{15.162}$$

The colons : ... : again denote normal ordering which, for fermionic fields, involves a change of sign when two operators are interchanged. For example, $: b_1 c_{-2} : = -c_{-2} b_1$ because (see below) c_{-2} is a creation operator and b_1 is an annihilation operator. The normal-ordering constant $-\delta_{n,0}$, which just adds -1 to $L_0^{(\mathrm{g})}$, is that needed to make $a^{(\mathrm{g})} = 0$ in the commutator (15.156). These generators have commutation relations with the expansion coefficients given by

$$[L_m^{(\mathrm{g})}, b_n] = (m-n) b_{m+n} \qquad [L_m^{(\mathrm{g})}, c_n] = -(2m+n) c_{m+n}. \tag{15.163}$$

The worldsheet Hamiltonian is again $H^{(\mathrm{g})} = L_0^{(\mathrm{g})} + \widetilde{L}_0^{(\mathrm{g})}$ so, setting $m = 0$ in (15.163), we see that the b_n and c_n are annihilation operators for $n > 0$ and creation operators for $n < 0$. The two operators b_0 and c_0, which commute with $L_0^{(\mathrm{g})}$, are neither creation nor annihilation operators. For the purposes of the definition of normal ordering in (15.162), c_0 is counted as a creation operator while b_0 is counted as an annihilation operator, but this is at present merely a matter of convention.

In the past, we have constructed a Hilbert space by identifying a unique ground state $|0\rangle$, which is annihilated by all the annihilation operators. In the present instance, this means $b_n|0\rangle = c_n|0\rangle = 0$ for all $n > 0$. Here, these conditions do not identify a unique ground state, because b_0 and c_0 are neither creation nor annihilation operators. In fact, there are two ghost ground states, $|0_\mathrm{g}\rangle$ and $|1_\mathrm{g}\rangle$, which are distinguished by the actions of b_0 and c_0, namely

$$b_0|0_\mathrm{g}\rangle = 0 \qquad b_0|1_\mathrm{g}\rangle = |0_\mathrm{g}\rangle \qquad c_0|0_\mathrm{g}\rangle = |1_\mathrm{g}\rangle \qquad c_0|1_\mathrm{g}\rangle = 0 \tag{15.164}$$

(see exercise 15.7). Thus, a basis for the ghost Hilbert space consists of two towers of states built on $|0_\mathrm{g}\rangle$ and $|1_\mathrm{g}\rangle$ by acting with arbitrarily many creation operators. Because the ghosts are fermions, however, we have $b_n^2 = c_n^2 = 0$ for all n, and each creation operator can act only once. This aside, the situation is analogous to the existence of an infinity of states $|0; k\rangle$ for the X^μ field theory, which are all ground states for the vibration modes but distinguished by the eigenvalues k^μ of the spacetime momentum $p^\mu = (2/\alpha')^{1/2}\alpha_0^\mu$. We shall eventually choose $|0_\mathrm{g}\rangle$ as the 'true' ground state, but at present this is merely a matter of notation.

15.3.5 The BRST cohomology

Let us again take stock of our position. We learned at the outset that the allowed states of the string are restricted by the constraints $T_{ab} = 0$. Therefore, not all of the basis vectors $\alpha_{n_1}^{\mu_1} \cdots \alpha_{n_N}^{\mu_N}|0; k\rangle$ can represent allowed, physically distinct states. From this point of view, the gauge fixing that we have taken so much trouble to set up might seem like a retrograde step, because the number of these basis vectors has been augmented by the presence of ghosts! Let us denote by $\mathcal{H}$ the Hilbert space that is spanned by all the basis vectors we have discovered. This basis consists of the ground states $|0; 0_\mathrm{g}; k\rangle$ and $|0; 1_\mathrm{g}; k\rangle$, where the first 0 denotes the ground state of the α_n^μ and $\widetilde{\alpha}_n^\mu$ oscillators, together with all the states that can be formed by acting with creation operators. (In the case of a closed string, there are also two possible ground states $|\widetilde{0}_\mathrm{g}\rangle$ and $|\widetilde{1}_\mathrm{g}\rangle$ for the independent left-moving ghosts.) Clearly, the Hilbert space that represents the physically allowed states of the string, say $\mathcal{H}_\mathrm{phys}$, must be much smaller than $\mathcal{H}$ and our task in this section is to construct it.

The key to this construction lies in a symmetry of the gauge-fixed action which, in terms of the coordinates z and $\bar{z}$ on the Euclidean worldsheet, is now

$$S = -\frac{1}{2\pi\alpha'}\int \mathrm{d}z\mathrm{d}\bar{z}\left[\partial X_\mu \bar{\partial} X^\mu - \alpha'\left(b\,\bar{\partial}c + \widetilde{b}\,\partial\widetilde{c}\right)\right]. \tag{15.165}$$

The *BRST transformation*, which leaves this action invariant, consists in changing the fields by

$$\delta X^\mu = \mathrm{i}\epsilon\left(c\,\partial X^\mu + \widetilde{c}\,\bar{\partial} X^\mu\right) \tag{15.166}$$

$$\delta b = -\mathrm{i}\epsilon\left(T^{(X)} + T^{(\mathrm{g})}\right) \tag{15.167}$$

$$\delta \widetilde{b} = -\mathrm{i}\epsilon\left(\widetilde{T}^{(X)} + \widetilde{T}^{(\mathrm{g})}\right) \tag{15.168}$$

$$\delta c = \mathrm{i}\epsilon\, c\,\partial c \tag{15.169}$$

$$\delta \widetilde{c} = \mathrm{i}\epsilon\, \widetilde{c}\,\bar{\partial}\widetilde{c} \tag{15.170}$$

where $T \equiv T_{zz}$ and $\widetilde{T} \equiv T_{\bar{z}\bar{z}}$. The parameter ϵ is an anticommuting constant, so terms of order ϵ^2 vanish and this is an exact, rather than merely an infinitesimal transformation. To be precise, the integrand in (15.165) changes by an amount

$$\mathrm{i}\epsilon\left[\partial\left(c\,\partial X_\mu \bar{\partial} X^\mu + \alpha' b\, c\,\bar{\partial} c\right) + \bar{\partial}\left(\widetilde{c}\,\partial X_\mu \bar{\partial} X^\mu + \alpha' \widetilde{b}\,\widetilde{c}\,\partial\widetilde{c}\right)\right] \tag{15.171}$$

which is a total divergence, so the action itself is invariant. This symmetry is valid at the classical level, where we treat the X^μ as real functions and b and c as anticommuting functions. It is a relic of the gauge invariance of the original theory. In fact, the transformation (15.166) of X^μ is just a coordinate transformation

$$X^\mu(z,\bar{z}) + \delta X^\mu(z,\bar{z}) = X^\mu(z + \mathrm{i}\epsilon c, \bar{z} + \mathrm{i}\epsilon\widetilde{c}). \tag{15.172}$$

A property that will prove crucial is that if we make a second BRST transformation with a different parameter ϵ', then $\delta'(\delta X^\mu) = 0$ and similarly for the ghost fields, provided that the equations of motion are satisfied. More explicitly, this means

$$\begin{aligned}\delta'(\delta X^\mu) &= \mathrm{i}\epsilon\delta'\left(c\,\partial X^\mu + \widetilde{c}\,\bar{\partial} X^\mu\right)\\ &= \mathrm{i}\epsilon\left[(\delta' c)\partial X^\mu + c\partial(\delta' X^\mu) + (\delta'\widetilde{c})\bar{\partial} X^\mu + \widetilde{c}\,\bar{\partial}(\delta' X^\mu)\right]\\ &= 0\end{aligned} \tag{15.173}$$

provided that $\bar{\partial} c = \partial\widetilde{c} = \partial\bar{\partial} X^\mu = 0$ and so on, as may easily be checked by substituting the explicit forms of $\delta' X^\mu, \ldots$ from (15.166)–(15.170).

Quantum-mechanically, we deal with this transformation in much the same way that we dealt with the supersymmetry transformation in §12.7.4, introducing a *BRST charge* Q, such that

$$\begin{gathered}[\mathrm{i}\epsilon Q, X^\mu] = \delta X^\mu\\ [\mathrm{i}\epsilon Q, b] = \delta b \qquad [\mathrm{i}\epsilon Q, \widetilde{b}\,] = \delta\widetilde{b} \qquad [\mathrm{i}\epsilon Q, c] = \delta c \qquad [\mathrm{i}\epsilon Q, \widetilde{c}\,] = \delta\widetilde{c}.\end{gathered} \tag{15.174}$$

The property that $\delta'(\delta X^\mu) = \ldots = 0$ implies that this charge is *nilpotent*, which means

$$Q^2 = 0. \tag{15.175}$$

On account of the anticommuting nature of ϵ, we can use the Laurent expansions (15.95), (15.96) and (15.161) to find that the commutators (15.174) are equivalent to

$$\begin{gathered}[Q, \alpha_n^\mu] = -n \sum_m c_m \alpha_{n-m}^\mu \\ \{Q, b_n\} = L_n^{(X)} + L_n^{(\mathrm{g})} \qquad \{Q, c_n\} = \tfrac{1}{2}\sum_m (2m-n) c_m c_{n-m}\end{gathered} \tag{15.176}$$

with similar relations for $\widetilde{\alpha}_n^\mu$, $\widetilde{b}_n$ and $\widetilde{c}_n$ in the case of a closed string. These determine the charge Q uniquely, and it can be expressed as

$$Q = \sum_{n=-\infty}^{\infty} \left[c_n L_{-n}^{(X)} + \widetilde{c}_n \widetilde{L}_{-n}^{(X)} + \tfrac{1}{2} : \left(c_n L_{-n}^{(\mathrm{g})} + \widetilde{c}_n \widetilde{L}_{-n}^{(\mathrm{g})} \right) : \right] - \tfrac{1}{2}(c_0 + \widetilde{c}_0) \tag{15.177}$$

as some patient algebra should serve to verify.

The use of the BRST symmetry in constructing the physical Hilbert space is roughly this. When we fixed the gauge, we made a more or less arbitrary decision to do this in such a way that the worldsheet metric became η_{ab}. This is certainly a great convenience, but in principle we could have extracted the gauge volume $\mathcal{V}_{\text{gauge}}$ by inserting some other fixed metric into the δ function in (15.130). Physical quantities, such as the probability amplitudes that we calculate from scalar products $\langle\psi'|\psi\rangle$ of the vectors in $\mathcal{H}_{\text{phys}}$ ought not to depend on this choice of metric. In particular, we may demand that an infinitesimal change in this choice of metric should leave these scalar products unchanged. The condition for this to be true is that $\mathrm{e}^{\mathrm{i}\epsilon Q}|\psi\rangle = |\psi\rangle$ for each vector in $\mathcal{H}_{\text{phys}}$, or that

$$Q|\psi\rangle = 0. \tag{15.178}$$

This assertion should, I hope, appear plausible, in view of the fact that the BRST symmetry is inherited from the original gauge invariance, but I propose to omit the wearisome details needed to prove it. Interested readers will find discussions in, for example, Green, Schwarz and Witten (1987) and Polchinski (1998).

It might seem that $\mathcal{H}_{\text{phys}}$ should consist of just those states in $\mathcal{H}$ for which $Q|\psi\rangle = 0$, but this is not quite good enough. In the language that we met briefly in §3.7 in connection with the exterior derivative d (which is also a nilpotent operator) a state that satisfies (15.178) can be called a *closed* state. (This has nothing to do with a 'closed' string.) There are some closed states of a special kind, namely those that can be expressed as $|\psi\rangle = Q|\chi\rangle$, where $|\chi\rangle$ may be any vector in $\mathcal{H}$. They are closed for the special reason that $Q|\psi\rangle = Q^2|\chi\rangle = 0$ and are called *exact* states. To make things work smoothly at this point, it is necessary that Q should be Hermitian, $Q^\dagger = Q$. This involves a technicality that I shall mention below, but suppose it is true. Then the bra vector corresponding to a closed ket vector $|\psi_{\rm closed}\rangle$ satisfies

$$\langle\psi_{\rm closed}|Q = 0 \tag{15.179}$$

and the bra vector corresponding to an exact ket vector $|\psi_{\rm exact}\rangle = Q|\chi\rangle$ is

$$\langle\psi_{\rm exact}| = \langle\chi|Q. \tag{15.180}$$

Now let $|\psi_1\rangle$ and $|\psi_2\rangle$ be any two closed vectors, which we hope to associate with physical states. From these, we can form two new vectors, by adding to them some arbitrary exact vectors, say

$$|\psi_1'\rangle = |\psi_1\rangle + Q|\chi_1\rangle \qquad |\psi_2'\rangle = |\psi_1\rangle + Q|\chi_2\rangle. \tag{15.181}$$

It is trivial to see that $|\psi_1'\rangle$ and $|\psi_2'\rangle$ are also closed and, furthermore, that

$$\langle\psi_2'|\psi_1'\rangle = \langle\psi_2|\psi_1\rangle. \tag{15.182}$$

This implies that $|\psi_1'\rangle$ and $|\psi_1\rangle$ (and similarly $|\psi_2'\rangle$ and $|\psi_2\rangle$) carry exactly the same physical information. The difference between them is accounted for by gauge degrees of freedom, which have no physical meaning. Thus, a physical state is represented not by a single vector in $\mathcal{H}$ but by a whole collection of vectors that differ from each other by the addition of arbitrary exact vectors.

We shall say that two vectors in $\mathcal{H}$ are *equivalent* if they differ only by an exact vector. A few moments thought (aided, perhaps, by the considerations of exercise 10.5) should enable readers to convince themselves that the set $\mathcal{H}_{\rm closed}$ of all the closed vectors in $\mathcal{H}$ can be split into *equivalence classes* such that all the vectors in one class are equivalent to each other, but no two vectors belonging to different classes are equivalent. *It is one of these equivalence classes that represents a definite physical state.* Now, these equivalence classes can themselves be regarded as vectors, say $|\Psi\rangle\rangle$, which form a Hilbert space. To make them into a Hilbert space, we simply need rules for adding vectors and forming scalar products, and these rules are ready to hand. Consider two equivalence classes, $|\Psi_1\rangle\rangle$ and $|\Psi_2\rangle\rangle$, and pick any one vector from each of them, say $|\psi_1\rangle$ and $|\psi_2\rangle$. The sum of these vectors, $|\psi_3\rangle = |\psi_1\rangle + |\psi_2\rangle$, belongs to some equivalence class $|\Psi_3\rangle\rangle$. Had we chosen any other pair of vectors, their sum would differ from $|\psi_3\rangle$ by some exact vector, and would also belong to $|\Psi_3\rangle\rangle$. We therefore have an unambiguous rule for the sum of equivalence classes:

$$|\Psi_1\rangle\rangle + |\Psi_2\rangle\rangle = |\Psi_3\rangle\rangle. \tag{15.183}$$

Similarly, we can define the scalar product

$$\langle\langle\Psi_2|\Psi_1\rangle\rangle = \langle\psi_2|\psi_1\rangle \tag{15.184}$$

which, on account of (15.182), does not depend on which representative vectors $|\psi_1\rangle$ and $|\psi_2\rangle$ we choose. The new Hilbert space constructed in this way is the *BRST cohomology* (or, more accurately, the cohomology of the BRST charge Q). I shall denote it by $\mathcal{H}_{\mathrm{BRST}}$. A rough and ready description is that we take the set of closed states $\mathcal{H}_{\mathrm{closed}}$ and 'divide out' the set of exact states $\mathcal{H}_{\mathrm{exact}}$ and this is reflected in the mathematical symbolism

$$\mathcal{H}_{\mathrm{BRST}} = \frac{\mathcal{H}_{\mathrm{closed}}}{\mathcal{H}_{\mathrm{exact}}}. \tag{15.185}$$

Finally, it may or may not be possible to interpret an operator A that acts in $\mathcal{H}$ as an operator that acts in $\mathcal{H}_{\mathrm{BRST}}$. Suppose that A acts on *any* exact vector $Q|\chi\rangle$ to produce another exact vector

$$AQ|\chi\rangle = Q|\chi'\rangle. \tag{15.186}$$

It follows that

$$QAQ|\chi\rangle = 0 \qquad \text{or} \qquad QAQ = 0 \tag{15.187}$$

because $|\chi\rangle$ can be any vector. Given an operator A with this property, suppose that it acts on a vector $|\psi_1\rangle$ belonging to the equivalence class $|\Psi_1\rangle\rangle$ to produce the vector

$$|\psi_2\rangle = A|\psi_1\rangle \tag{15.188}$$

and that $|\psi_2\rangle$ belongs to the equivalence class $|\Psi_2\rangle\rangle$. We can then say that

$$A|\Psi_1\rangle\rangle = |\Psi_2\rangle\rangle \tag{15.189}$$

because, on account of (15.186), the action of A on some other vector, say $|\psi_1\rangle + Q|\chi\rangle$, belonging to $|\Psi_1\rangle\rangle$ produces the vector

$$A\Big(|\psi_1\rangle + Q|\chi\rangle\Big) = |\psi_2\rangle + Q|\chi'\rangle \tag{15.190}$$

which also belongs to $|\Psi_2\rangle\rangle$. We can call A a *gauge invariant* operator if it has the property (15.187), because it has a physical meaning, expressed by (15.189), independent of the gauge degrees of freedom contained in the exact vectors. Thus, a gauge-invariant operator in $\mathcal{H}$ can be interpreted as an operator in $\mathcal{H}_{\mathrm{BRST}}$. This new Hilbert space $\mathcal{H}_{\mathrm{BRST}}$ is almost, but not quite, the physical Hilbert space $\mathcal{H}_{\mathrm{phys}}$ that we hoped to construct.

We have yet to take account of the existence of the two ghost ground states $|0_{\mathrm{g}}\rangle$ and $|1_{\mathrm{g}}\rangle$ in (15.164). It should be fairly plausible that the physically relevant ground state is $|0_{\mathrm{g}}\rangle$, which obeys $b_0|0_{\mathrm{g}}\rangle = 0$, for the following reason. Taking into account all the other degrees of freedom, physical states will be those which obey the two conditions

$$Q|\psi\rangle = 0 \qquad \text{and} \qquad b_0|\psi\rangle = 0. \tag{15.191}$$

For a closed string, there will be a third condition $\widetilde{b}_0|\psi\rangle = 0$. The second of the equations (15.176) which define the charge Q then tells us that

$$\left[L_0^{(X)} + L_0^{(\mathrm{g})}\right]|\psi\rangle = \{Q, b_0\}|\psi\rangle = 0. \tag{15.192}$$

This is one of our original constraints $L_n = \widetilde{L}_n = 0$ except that, in its final gauge-fixed form, it includes the contribution of the ghosts. Finally, then, the physical Hilbert space $\mathcal{H}_{\mathrm{phys}}$ is the one that we obtain from $\mathcal{H}_{\mathrm{BRST}}$ by imposing the additional constraint $b_0|\psi\rangle = 0$.

In the interest of accuracy, I am now going to discuss a technicality that we deferred earlier on, namely the fact that the BRST charge Q must be Hermitian. The results of this discussion, though important, will not bear directly on what I have to say later on, so less fastidious readers may wish to skip the remainder of this section. The condition for Q to be Hermitian is straightforward, if tedious, to find. It is that the expansion coefficients for both the X^μ and the ghost fields must satisfy $\alpha_n^{\mu\dagger} = \alpha_{-n}^\mu$ and so on. In the case of the X^μ, we have already seen that this is the condition for these coordinates to be real; it also has the satisfactory consequence that α_n^μ and $\alpha_n^{\mu\dagger}$ are respectively the annihilation and creation operators for quanta of energy in the nth mode of vibration. What is disconcerting is that we also require b_0 and c_0 to be Hermitian: $b_0^\dagger = b_0$ and $c_0^\dagger = c_0$. Now, the matrices that represent these operators in exercise 15.7 are not, in the ordinary sense, Hermitian matrices. Whether they count as Hermitian operators in the Hilbert space $\mathcal{H}$ depends on our definition of the scalar product, though this is a matter that we have not previously needed to consider in detail. The definition of the scalar product also affects what we mean by a dual vector, just as a metric defines a correspondence between vectors and one-forms by raising and lowering of indices. In the present instance, if b_0 and c_0 are to be Hermitian, then the vectors $|0_{\mathrm{g}}\rangle = \binom{1}{0}$ and $|1_{\mathrm{g}}\rangle = \binom{0}{1}$ cannot be orthonormal, which means that the 'metric', say $\mathfrak{g}$, in this 2-dimensional space is not diagonal in this basis; in fact we must take

$$\mathfrak{g} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{15.193}$$

Using this matrix to 'raise the indices', we find that the basis bra vectors are

$$\langle 0_{\mathrm{g}}| = (\,1 \quad 0\,)\,\mathfrak{g} = (\,0 \quad 1\,) \qquad \langle 1_{\mathrm{g}}| = (\,0 \quad 1\,)\,\mathfrak{g} = (\,1 \quad 0\,)\,. \tag{15.194}$$

The matrices formed, as it were, by the matrix elements of b_0 and c_0 are

$$\begin{pmatrix} \langle 0_{\mathrm{g}}|b_0|0_{\mathrm{g}}\rangle & \langle 0_{\mathrm{g}}|b_0|1_{\mathrm{g}}\rangle \\ \langle 1_{\mathrm{g}}|b_0|0_{\mathrm{g}}\rangle & \langle 1_{\mathrm{g}}|b_0|1_{\mathrm{g}}\rangle \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{15.195}$$

$$\begin{pmatrix} \langle 0_{\mathrm{g}}|c_0|0_{\mathrm{g}}\rangle & \langle 0_{\mathrm{g}}|c_0|1_{\mathrm{g}}\rangle \\ \langle 1_{\mathrm{g}}|c_0|0_{\mathrm{g}}\rangle & \langle 1_{\mathrm{g}}|c_0|1_{\mathrm{g}}\rangle \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{15.196}$$

and these *are* Hermitian, which is what we need. Unfortunately, our ground-state vector now has zero length: $\langle 0_{\mathrm{g}}|0_{\mathrm{g}}\rangle = 0$. The only non-zero matrix element we can form using $|0_{\mathrm{g}}\rangle$ is $\langle 0_{\mathrm{g}}|c_0|0_{\mathrm{g}}\rangle = 1$. Including all the other degrees of freedom (represented, say, by ϕ), the non-zero matrix elements that we can construct after imposing the constraint $b_0|\psi\rangle = 0$ are of the form

$$\langle \phi_2; 0_{\mathrm{g}}|c_0|0_{\mathrm{g}}; \phi_1\rangle = \langle \phi_2|\phi_1\rangle \tag{15.197}$$

and it is this expression that must be used to construct the scalar product in $\mathcal{H}_{\mathrm{phys}}$. In the case of a closed string, of course, the left-moving ghost operators $\widetilde{b}_0$ and $\widetilde{c}_0$ must be treated in the same way.

15.4 Physics of the Free Bosonic String

We are finally in a position to extract from the formalism of the preceding sections some concrete conclusions about the physical properties of the quantized string. Of course, we are rather far from being able to identify this object with anything that is actually observed in nature. As it stands, the theory applies to a string that exists in a 26-dimensional spacetime and undergoes no interactions—and we shall soon find that another misfortune awaits us! The key questions that I plan to address in this section are, first, how we can interpret a string in a given state of vibration as a particle of definite mass and spin in spacetime and, second, how it is that string theory promises us a quantum theory of gravity.

15.4.1 The mass spectrum

I shall deal explicitly with the lowest-lying states of a closed string, and it will be useful to start by assembling the essential information that we have to work with. First of all, we have to study those states in the Hilbert space $\mathcal{H}$ which satisfy the conditions $Q|\psi\rangle = 0$ and $b_0|\psi\rangle = \widetilde{b}_0|\psi\rangle = 0$. We learned in (15.192) that these also imply $L_0|\psi\rangle = 0$ and, for the closed string, $\widetilde{L}_0|\psi\rangle = 0$. Here, L_0 and $\widetilde{L}_0$ are the Virasoro generators for the combined field theory of the X^μ and the ghosts. From the explicit expressions (15.103) with $a = 0$ and (15.162), we have

$$L_0 = -\frac{\alpha'}{4}M^2 + \sum_{n=1}^{\infty} n\left(N_n^{(X)} + N_n^{(b)} + N_n^{(c)}\right) - 1 \tag{15.198}$$

where

$$N_n^{(X)} = -n^{-1}\alpha_{-n\,\mu}\alpha_n^\mu \qquad N_n^{(b)} = b_{-n}c_n \qquad N_n^{(c)} = c_{-n}b_n. \tag{15.199}$$

These three operators count the numbers of quanta of energy in the nth vibrational modes of the X^μ and ghost fields (see exercise 15.8); the quantum of energy in the nth mode is proportional to n in each case. For the ghost modes, the quanta counted by $N_n^{(b)}$ are created by b_{-n} and annihilated by c_n, while the converse is true for the quanta counted by $N_n^{(c)}$. As in the classical formula (15.79), $M^2 = p_\mu p^\mu$ represents the mass^2 of the string, and for physical states, which obey the constraint $L_0|\psi\rangle = 0$, we can identify the mass^2 operator as

$$M^2 = \frac{4}{\alpha'}\left[\sum_{n=1}^{\infty} n\left(N_n^{(X)} + N_n^{(b)} + N_n^{(c)}\right) - 1\right]. \tag{15.200}$$

For the left-moving modes, we can write an exactly similar set of equations. In particular, the constraint $\widetilde{L}_0|\psi\rangle = 0$ tells us, as for the classical string, that the mass^2 is also given by

$$M^2 = \frac{4}{\alpha'}\left[\sum_{n=1}^{\infty} n\left(\widetilde{N}_n^{(X)} + \widetilde{N}_n^{(b)} + \widetilde{N}_n^{(c)}\right) - 1\right]. \tag{15.201}$$

As we might have expected, the vibrational states of the string correspond to a sequence of ‘energy levels’, labelled by an integer

$$N = \sum_{n=1}^{\infty} n\left(N_n^{(X)} + N_n^{(b)} + N_n^{(c)}\right) = \sum_{n=1}^{\infty} n\left(\widetilde{N}_n^{(X)} + \widetilde{N}_n^{(b)} + \widetilde{N}_n^{(c)}\right) \tag{15.202}$$

in terms of which we have $M^2 = (4/\alpha')(N-1)$.

These levels are degenerate; that is, each level corresponds in general to more than one state. To find out just how many physical states there are at each level, we need to take account of the BRST condition $Q|\psi\rangle = 0$ and of the equivalence classes discussed in the last section. For this purpose, the expression (15.177) can be rearranged to read

$$\begin{aligned} Q = &-\tfrac{1}{2}\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\sum_{n=-\infty}^{\infty}\left(c_m\alpha_{n\,\mu}\alpha^{\mu}_{-m-n} + \widetilde{c}_m\widetilde{\alpha}_{n\,\mu}\widetilde{\alpha}^{\mu}_{-m-n}\right)\\ &+\tfrac{1}{2}\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}(m-n)\left(:c_m c_n b_{-m-n}: + :\widetilde{c}_m\widetilde{c}_n\widetilde{b}_{-m-n}:\right)\\ &+c_0L_0 + \widetilde{c}_0\widetilde{L}_0 \end{aligned} \tag{15.203}$$

which is useful for two reasons. First, our physical states are supposed to obey both $b_0|\psi\rangle = 0$ and $Q|\psi\rangle = 0$. Now, it is easy to see from (15.164) that if $|\psi\rangle$ obeys $b_0|\psi\rangle = 0$ then the vector $c_0|\psi\rangle$ does *not* obey this constraint. It would be awkward, then, if Q were to contain the operator c_0 (or, for the same reason, $\widetilde{c}_0$). Fortunately, we see from (15.203) that the only terms in Q which do contain c_0 and $\widetilde{c}_0$ are also proportional to L_0 or $\widetilde{L}_0$. We can begin our construction of the physical Hilbert space $\mathcal{H}_{\text{phys}}$ by restricting our attention to the space $\mathcal{H}_0$ of vectors for which

$$b_0|\psi\rangle = L_0|\psi\rangle = \widetilde{L}_0|\psi\rangle = 0. \tag{15.204}$$

When Q acts on vectors in this space, the last two terms in (15.203) can be ignored. In that case, the (anti)commutators (15.176) can be written as

$$[Q, \alpha_n^\mu] = -n \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} c_m \alpha_{n-m}^\mu \tag{15.205}$$

$$\{Q, b_n\} = L_n^{(X)} + \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} (m+n) : b_{n-m} c_m : \tag{15.206}$$

$$\{Q, c_n\} = \tfrac{1}{2} \sum_{\substack{m=-\infty \\ m\neq 0,n}}^{\infty} (2m-n) c_m c_{n-m} \tag{15.207}$$

provided that all the operators are taken to act in the space $\mathcal{H}_0$. In particular, (15.206) and (15.207) hold for $n \neq 0$ and the operators b_0 and c_0 can be ignored entirely. The same applies, of course, to the left-moving modes.

The second useful feature of (15.203) is this. We learned in the last section that two states are physically equivalent if they differ by an exact vector, of the form $Q|\chi\rangle$. By looking at the combinations of creation and annihilation operators that appear in (15.203), it is not hard to see that $Q|\chi\rangle$ belongs to the same level as $|\chi\rangle$. We can therefore determine what the physically distinct states are by dealing with one level at a time.

A sensible place to start, perhaps, is the lowest level, $N = 0$. For a given spacetime momentum k^μ, there is one state, namely the ground state of all the oscillators. I denote this state by $|0; \boldsymbol{k}\rangle$, where 0 means the oscillator ground state and the 25-component vector $\boldsymbol{k}$ represents the spatial components (relative to some chosen frame of reference) of k^μ. The spacetime energy $k^0 = \sqrt{|\boldsymbol{k}|^2 + M^2}$ is determined, for every state, by the mass formulae (15.200) and (15.201), which express the constraints $L_0|\psi\rangle = \widetilde{L}_0|\psi\rangle = 0$. It is easy to see from (15.203) that the BRST condition $Q|0; \boldsymbol{k}\rangle$ is satisfied, because each term contains at least one annihilation operator. Since there are no other states at this level, there are no exact states and $|0; \boldsymbol{k}\rangle$ is an equivalence class in itself. The mass is given by $M^2 = -4/\alpha'$. *This is a disaster!* In classical terms, the relation $M^2 = E^2(1-v^2)$ shows that a particle with $M^2 < 0$ travels at a speed v greater than that of light. It is a *tachyon* which, as we saw in exercise 2.4, is inconsistent with the requirement of causality, that a cause should precede its effect.

Evidently, the bosonic string does not, in itself, provide a useful model for real relativistic particles. Nevertheless, some further investigation will be worthwhile, because it will reveal features that can be carried over to more sophisticated versions of string theory. Let us examine the states that arise at the level $N = 1$, for which $M^2 = (4/\alpha')(N-1) = 0$; these correspond to massless particles. We can create a state with $N = 1$ by acting on $|0; \boldsymbol{k}\rangle$ with any of the creation operators α_{-1}^μ, b_{-1} and c_{-1}. Because both (15.200) and (15.201) must hold simultaneously, we must act at the same time with one of the left-moving creation operators $\widetilde{\alpha}^{\mu}_{-1}$, $\widetilde{b}_{-1}$ and $\widetilde{c}_{-1}$. A general level-1 state can be expressed as

$$|\mathcal{O}_1; \boldsymbol{k}\rangle = A_{-1}(\epsilon, \kappa, \widetilde{\kappa}, \ldots)|0; \boldsymbol{k}\rangle \tag{15.208}$$

the creation operator being given by

$$A_{-1}(\epsilon, \kappa, \widetilde{\kappa}, \ldots) = \epsilon_{\mu\nu}\, \alpha^{\mu}_{-1}\widetilde{\alpha}^{\nu}_{-1} + \kappa_{\mu}\, \alpha^{\mu}_{-1}\widetilde{b}_{-1} + \widetilde{\kappa}_{\mu}\, \widetilde{\alpha}^{\mu}_{-1}b_{-1} + \ldots \tag{15.209}$$

where $\epsilon_{\mu\nu}$, κ_{μ}, $\widetilde{\kappa}_{\mu}$, ... are constants and '...' represents all the other possible terms involving one right-moving and one left-moving operator. The algebra needed to extract the equivalence class of physical states is straightforward, but rather cumbersome. I shall indicate how it works, and leave sufficiently energetic readers to fill in the details. (The corresponding algebra for an open string is much easier, and readers are invited to tackle it in exercise 15.9.) First, let us act on the general level-1 vector with Q. Using the (anti)commutators (15.205)–(15.207) and taking into account that $Q|0; \boldsymbol{k}\rangle = 0$ and $\alpha^{\mu}_{0}|0; \boldsymbol{k}\rangle = (\alpha'/2)^{1/2}k^{\mu}|0; \boldsymbol{k}\rangle$, we find

$$Q|\mathcal{O}_1; \boldsymbol{k}\rangle = \left(\frac{\alpha'}{2}\right)^{1/2} B_{-1}(\epsilon, \ldots; k)|0; \boldsymbol{k}\rangle \tag{15.210}$$

$$\begin{aligned} B_{-1}(\epsilon, \ldots; k) = \Big[&\epsilon_{\mu\nu}\left(k^{\mu}c_{-1}\widetilde{\alpha}^{\nu}_{-1} + k^{\nu}\alpha^{\mu}_{-1}\widetilde{c}_{-1}\right) - \left(\kappa_{\mu}k_{\nu} + k_{\mu}\widetilde{\kappa}_{\nu}\right)\alpha^{\mu}_{-1}\widetilde{\alpha}^{\nu}_{-1} \\ &-k^{\mu}\left(\kappa_{\mu}c_{-1}\widetilde{b}_{-1} + \widetilde{\kappa}_{\mu}\tilde{c}_{-1}b_{-1}\right) + \ldots\Big]. \end{aligned} \tag{15.211}$$

From this we can deduce two things. By setting $B(\epsilon, \ldots; k) = 0$, we get a set of conditions on the coefficients $\epsilon_{\mu\nu}$, ... which, if they are satisfied, will make $|\mathcal{O}_1; \boldsymbol{k}\rangle$ a closed vector. On the other hand, (15.210) is itself the general form of an exact vector at level 1, and any two closed vectors which differ by a vector of this form are equivalent, in the sense we discussed in the previous section. The upshot is that any closed state is equivalent to a state of the form

$$|\mathcal{O}^{\text{closed}}_1; \boldsymbol{k}\rangle = \epsilon_{\mu\nu}(k)\alpha^{\mu}_{-1}\widetilde{\alpha}^{\nu}_{-1}|0; \boldsymbol{k}\rangle \tag{15.212}$$

where $\epsilon_{\mu\nu}$ obeys the 'transversality' condition

$$k^{\mu}\epsilon_{\mu\nu}(k) = k^{\nu}\epsilon_{\mu\nu}(k) = 0 \tag{15.213}$$

and two states with polarization tensors $\epsilon_{\mu\nu}(k)$ and $\epsilon'_{\mu\nu}(k)$ are equivalent if

$$\epsilon'_{\mu\nu}(k) = \epsilon_{\mu\nu}(k) + \kappa_{\mu}k_{\nu} + k_{\mu}\widetilde{\kappa}_{\nu} \tag{15.214}$$

where κ_{μ} and $\widetilde{\kappa}_{\nu}$ are any spacetime vectors such that

$$k^{\mu}\kappa_{\mu} = k^{\mu}\widetilde{\kappa}_{\mu} = 0. \tag{15.215}$$

Note that, as we might have hoped, the operators that create ghost excitations do not figure in (15.212). Indeed, their net effect is to *reduce* the number of physical degrees of freedom through the gauge equivalence (15.214). The same is true at all levels of excitation—a statement which goes under the name of the 'no-ghost theorem'.

To appreciate the implications of this result in a simple way, let us imagine that the number of spacetime dimensions is $d = 4$; the true level-1 content of the bosonic string theory is a generalization to 26 dimensions of what we shall find out in this way. The polarization tensor $\epsilon^{\mu\nu}$ can be split up as

$$\epsilon^{\mu\nu} = \epsilon_{\mathrm{a}}^{\mu\nu} + \epsilon_{\mathrm{g}}^{\mu\nu} + \phi\eta^{\mu\nu} \tag{15.216}$$

where the three constituent tensors are

$$\epsilon_{\mathrm{a}}^{\mu\nu} = \tfrac{1}{2}\left(\epsilon^{\mu\nu} - \epsilon^{\nu\mu}\right) \tag{15.217}$$

$$\epsilon_{\mathrm{g}}^{\mu\nu} = \tfrac{1}{2}\left(\epsilon^{\mu\nu} + \epsilon^{\nu\mu}\right) - \tfrac{1}{4}\epsilon_{\lambda}^{\lambda}\eta^{\mu\nu} \tag{15.218}$$

$$\phi = \tfrac{1}{4}\epsilon_{\lambda}^{\lambda}. \tag{15.219}$$

Of these, $\epsilon_{\mathrm{a}}^{\mu\nu}$ is an antisymmetric rank-2 tensor; $\epsilon_{\mathrm{g}}^{\mu\nu}$ is a symmetric rank-2 tensor which is traceless ($\epsilon_{\mathrm{g}\,\mu}^{\mu} = 0$) and ϕ is a scalar. The rationale for this decomposition is that each of the three constituent tensors forms an 'irreducible representation' of the Poincaré group, which is to say that each of them transforms separately under Lorentz transformations and rotations, and that they cannot be split further into tensors that have this property. We regard each of them as corresponding to a particle of definite spin.

Consider first the antisymmetric tensor $\epsilon_{\mathrm{a}}^{\mu\nu}$. In four dimensions, a general antisymmetric tensor has six independent components. However, the transversality constraint (15.213) and the equivalence (15.214) actually imply that $\epsilon_{\mathrm{a}}^{\mu\nu}$ contains only one physical degree of freedom, corresponding to a spin-0 particle called an *axion* (see exercise 15.10).

The symmetric, traceless tensor $\epsilon_{\mathrm{g}}^{\mu\nu}$ has a more fundamental significance. According to (15.214), it is physically equivalent to another tensor

$$\epsilon_{\mathrm{g}}^{\prime\mu\nu} - k^{\mu}\theta^{\nu} - k^{\nu}\theta^{\mu} \tag{15.220}$$

where $\theta^{\mu} = -\frac{1}{2}\left(\kappa^{\mu} + \widetilde{\kappa}^{\mu}\right)$ is an arbitrary vector with the property $k_{\mu}\theta^{\mu} = 0$. This is just the same as the gauge ambiguity (7.127) that we met in connection with the polarization tensor of a *graviton* (which is also a symmetric rank-2 tensor), apart from the following detail. The $\epsilon^{\mu\nu}$ appearing in (7.127) does not satisfy the transversality condition $k_{\mu}\epsilon^{\mu\nu} = k_{\nu}\epsilon^{\mu\nu} = 0$ nor does its trace vanish. This is offset, however, by the fact that θ^{μ} in (7.127) is not transverse either. In fact, by choosing $k_{\mu}\theta^{\mu} = \frac{1}{2}\epsilon_{\mu}^{\mu}$ in (7.127), we can make $\bar{\epsilon}^{\mu\nu}$ traceless, and (7.125) then shows that it will also be transverse. Once we have done this, the remaining gauge freedom corresponds exactly to (15.220), with $k_{\mu}\theta^{\mu} = 0$. This shows us, at least, that the particle corresponding to $\epsilon_{\mathrm{g}}^{\mu\nu}$ is (in four dimensions) a spin-2 particle like the graviton. Whether it has the same interpretation in terms of spacetime geometry and gravitational forces is another matter, which we shall investigate in the next two subsections.

The remaining massless degree of freedom, the scalar ϕ, must correspond to another spin-0 particle. It is called the *dilaton*, for the following reason. Suppose that in (7.124) we choose $\epsilon^{\mu\nu} = \phi\eta^{\mu\nu}$. Then the perturbed metric is

$$\eta^{\mu\nu} + h^{\mu\nu} = \eta^{\mu\nu}\left[1 + \phi_k(x)\right] \approx \exp\left[\phi_k(x)\right]\eta^{\mu\nu} \tag{15.221}$$

with $\phi_k(x) = \phi \mathrm{e}^{-\mathrm{i}k\cdot x}$ and this is a Weyl rescaling or dilation such as we have had much cause to think about in this chapter. Note carefully, though, that it is the spacetime metric rather than the worldsheet metric that presently concerns us. This dilaton is therefore quite different from the one I mentioned briefly in connection with non-critical strings (§15.3.3). The name is also something of a misnomer. It should be clear from the discussion of the previous paragraph that ϕ is in fact a gauge degree of freedom as far as general relativity is concerned. In the string theory, we learn from (15.214) that

$$\phi' = \epsilon'^{\mu}_{\mu} = \epsilon^{\mu}_{\mu} = \phi \tag{15.222}$$

on account of the transversality $k_\mu\kappa^\mu = k_\mu\widetilde{\kappa}^\mu = 0$, so ϕ cannot be set to zero by any choice of κ and $\widetilde{\kappa}$ and is *not* a gauge degree of freedom. It cannot, therefore be straightforwardly identified with a spacetime Weyl transformation.

15.4.2 Vertex operators

In (15.108), we defined a state $|0; k\rangle$, which is annihilated by all the α_n^μ for $n \geq 1$ (and would also have been annihilated by the $\widetilde{\alpha}_n^\mu$ for $n \geq 1$ had we been taking account of them at that point). It is also an eigenstate of the spacetime momentum operators p^μ, with eigenvalues k^μ. This state must be carefully distinguished from the tachyon state $|0; \boldsymbol{k}\rangle$, which has independent eigenvalues for the *spatial* momenta, $\boldsymbol{p}|0; \boldsymbol{k}\rangle = \boldsymbol{k}|0; \boldsymbol{k}\rangle$, but which yields a value for p^0 determined by the constraint $L_0|0; \boldsymbol{k}\rangle = 0$. I will now denote by $|\Omega\rangle$ the state that we get by adding to $|0; k\rangle$ the ghost ground state and setting $k^\mu = 0$, for $\mu = 0, \ldots, 25$. In particular, it is annihilated by the p^μ:

$$|\Omega\rangle = |0; 0_{\mathrm{g}}; k^\mu = 0\rangle \qquad p^\mu|\Omega\rangle = 0. \tag{15.223}$$

This state is uniquely defined, because a Lorentz transformation of the 26-vector $k^\mu = 0$ does not change it. In fact, $|\Omega\rangle$ can usefully be regarded as the overall ground state of the entire Hilbert space $\mathcal{H}$; we can generate a complete set of basis vectors from it by acting with the creation operators and with the centre of mass coordinates x^μ, which we have not used until now. However, $|\Omega\rangle$ does not belong to one of the equivalence classes from which we constructed the physical Hilbert space $\mathcal{H}_{\mathrm{phys}}$, because it does not satisfy the constraint $L_0|\psi\rangle = 0$. In fact, we see from (15.198) that $L_0|\Omega\rangle = \widetilde{L}_0|\Omega\rangle = -|\Omega\rangle$, so $|\Omega\rangle$ does not represent a physical state of the string.

It turns out to be extremely useful to find the operators which produce physical states from $|\Omega\rangle$ (or, at least, representative vectors in the equivalence classes that we regard as the physical states). Let us say that

$$\mathcal{V}(\mathcal{O};\boldsymbol{k})|\Omega\rangle = |\mathcal{O};\boldsymbol{k}\rangle \tag{15.224}$$

where $\mathcal{O}$ again represents the state of the vibration modes. The operator $\mathcal{V}(\mathcal{O};\boldsymbol{k})$ is called a *vertex operator*. As a matter of fact, if we can find the vertex operator $\mathcal{V}_{\mathrm{t}}(\boldsymbol{k}) = \mathcal{V}(0;\boldsymbol{k})$ for the tachyon state, then we can find all the states at higher levels by acting with the creation operators α^{μ}_{-n} and $\tilde{\alpha}^{\mu}_{-n}$. Now, the tachyon state $|0;\boldsymbol{k}\rangle$ differs from $|\Omega\rangle$ only by the eigenvalues of p^{μ}, and a simple modification of the results of exercise 5.3 tells us how to change these eigenvalues. In fact, taking into account the commutation relation $[p^{\mu}, x^{\nu}] = \mathrm{i}\eta^{\mu\nu}$, we find

$$p^{\mu}\mathrm{e}^{-\mathrm{i}k\cdot x} = \mathrm{e}^{-\mathrm{i}k\cdot x}\left(p^{\mu}+k^{\mu}\right) \qquad \text{and so} \qquad p^{\mu}\mathrm{e}^{-\mathrm{i}k\cdot x}|\Omega\rangle = k^{\mu}\mathrm{e}^{-\mathrm{i}k\cdot x}|\Omega\rangle \tag{15.225}$$

with $k\cdot x = k_{\mu}x^{\mu}$. Up to a possible normalization constant, therefore, the tachyon vertex operator is given by the simple expression

$$\mathcal{V}_{\mathrm{t}}(\boldsymbol{k}) \simeq \mathrm{e}^{-\mathrm{i}k\cdot x} \tag{15.226}$$

on the understanding that $k^0 = \left(|\boldsymbol{k}|^2 + M_{\mathrm{t}}^2\right)^{1/2}$, where $M_{\mathrm{t}}^2 = -4/\alpha'$ is the mass^2 of the (closed string) tachyon.

As occasionally happens in string theory, there is some virtue in making this simple matter more complicated. To be more honest, the $\simeq$ indicates that this is not really the definition of the vertex operator—it is a prototype of a more complicated object that is central to a theory of interacting strings. To construct the real vertex operator, I shall first split the mode expansion (15.95) for X^{μ} into two parts, separating the creation and annihilation operators:

$$X^{\mu}_{\mathrm{an}}(z,\bar{z}) = -\tfrac{1}{2}\mathrm{i}\alpha' p^{\mu}\ln(\bar{z}z) + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2}\sum_{n=1}^{\infty}\frac{1}{n}\left[\alpha^{\mu}_{n}z^{-n} + \tilde{\alpha}^{\mu}_{n}\bar{z}^{-n}\right] \tag{15.227}$$

$$X^{\mu}_{\mathrm{cr}}(z,\bar{z}) = x^{\mu} - \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2}\sum_{n=1}^{\infty}\frac{1}{n}\left[\alpha^{\mu}_{-n}z^{n} + \tilde{\alpha}^{\mu}_{-n}\bar{z}^{n}\right]. \tag{15.228}$$

On account of (15.223), I have grouped the momenta p^{μ} with the annihilation operators and their conjugate operators x^{μ} with the creation operators. With this convention, we can define a normal-ordered operator

$$\begin{aligned}\mathcal{V}_{\mathrm{t}}(z,\bar{z};\boldsymbol{k}) &= :\exp\left[-\mathrm{i}k\cdot X(z,\bar{z})\right]: \\ &= \exp\left[-\mathrm{i}k\cdot X_{\mathrm{cr}}(z,\bar{z})\right]\exp\left[-\mathrm{i}k\cdot X_{\mathrm{an}}(z,\bar{z})\right].\end{aligned} \tag{15.229}$$

Because p^{μ} commutes with everything except x^{μ}, it is still the case that

$$p^{\mu}\mathcal{V}_{\mathrm{t}}(z,\bar{z};\boldsymbol{k}) = \mathcal{V}_{\mathrm{t}}(z,\bar{z};\boldsymbol{k})\left(p^{\mu}+k^{\mu}\right). \tag{15.230}$$

Consider the effect of this operator on $|\Omega\rangle$. Acting with the rightmost exponential has no effect, because every operator in the exponent gives zero, so we have

$$\mathcal{V}_{\mathrm{t}}(z,\bar{z};\boldsymbol{k})|\Omega\rangle = \exp\left[-\mathrm{i}k\cdot X_{\mathrm{cr}}(z,\bar{z})\right]|\Omega\rangle. \tag{15.231}$$

In the remaining exponential, all the α^{μ}_{-n} and $\widetilde{\alpha}^{\mu}_{-n}$ are multiplied by positive powers of z or $\bar{z}$, so we find

$$\mathcal{V}_{\mathrm{t}}(0,0;\boldsymbol{k})|\Omega\rangle = \mathrm{e}^{-\mathrm{i}k\cdot x}|\Omega\rangle = |0;\boldsymbol{k}\rangle. \tag{15.232}$$

Thus, although the vertex operator $\mathcal{V}_{\mathrm{t}}(0,0;\boldsymbol{k})$ is by no means the same as its prototype (15.226), it has exactly the same effect on the ground state $|\Omega\rangle$.

I shall be a little more explicit in the next section about the use of vertex operators in a theory of interacting strings. Here, let us just observe that $\mathcal{V}_{\mathrm{t}}(0,0;\boldsymbol{k})$ is a local operator on the world sheet: it acts at the point $z=0$ which, according to figure 15.2(*b*) is equivalent to the $\tau \to -\infty$ end of the cylindrical worldsheet of figure 15.2(*a*). It may seem plausible, then, that by acting on $|\Omega\rangle$ with more vertex operators at other points, we could create a worldsheet with more than two ends, and that this might represent processes such as the merging, emission or absorption of several strings.

Of more immediate interest is the task of constructing the vertex operator for a graviton, which I shall use in the next subsection to show that we really do have the possibility of a quantum theory of gravity. Fairly obviously, the prototype vertex operator corresponding to (15.226) for a level-1 state (15.212) is

$$\mathcal{V}_{\mathrm{g}}(\boldsymbol{k}) \simeq \epsilon_{\mu\nu}(k)\alpha^{\mu}_{-1}\widetilde{\alpha}^{\nu}_{-1}\mathrm{e}^{-\mathrm{i}k\cdot x} \tag{15.233}$$

with, in this case, $k^0 = |\boldsymbol{k}|$ and the state will be a graviton if the polarization tensor $\epsilon_{\mu\nu}$ is chosen appropriately. How to construct the complete vertex operator becomes clear from considering the derivatives of X^{μ}_{cr} which, from (15.228), are

$$\partial X^{\mu}_{\mathrm{cr}}(z) = -\mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2}\left[\alpha^{\mu}_{-1} + \sum_{n=1}^{\infty}\alpha^{\mu}_{-n-1}z^{n}\right] \tag{15.234}$$

$$\bar{\partial} X^{\mu}_{\mathrm{cr}}(\bar{z}) = -\mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2}\left[\widetilde{\alpha}^{\mu}_{-1} + \sum_{n=1}^{\infty}\widetilde{\alpha}^{\mu}_{-n-1}\bar{z}^{n}\right]. \tag{15.235}$$

When $z=\bar{z}=0$, the only surviving terms are the creation operators that we need in (15.233), and a little thought will show that the correct expression is

$$\mathcal{V}_{\mathrm{g}}(z,\bar{z};\boldsymbol{k}) = -\frac{2}{\alpha'}\epsilon_{\mu\nu} : \partial X^{\mu}(z)\bar{\partial}X^{\nu}(\bar{z})\exp\left[-\mathrm{i}k\cdot X(z,\bar{z})\right] : . \tag{15.236}$$

Acting with this operator on $|\Omega\rangle$, we see that the normal ordering makes all the annihilation operators act first, giving zero; the only surviving terms consist purely of creation operators and after setting $z=\bar{z}=0$ in these terms we get the prototype vertex (15.233).

15.4.3 Strings and quantum gravity

So far, we have thought about a quantum-mechanical string which propagates through Minkowski spacetime. If the spacetime is curved, say with a metric $g_{\mu\nu}(x)$, then by analogy with (4.2) the string action on the Euclidean worldsheet must be

$$S_{\mathrm{E}} = -\frac{1}{2\pi\alpha'}\int \mathrm{d}z\mathrm{d}\bar{z}\, g_{\mu\nu}(X)\partial X^{\mu}\bar{\partial}X^{\nu}. \tag{15.237}$$

This defines a more complicated theory, because it is no longer quadratic in the fields X^{μ}. Considered as a two-dimensional field theory, it now contains interactions—the non-quadratic terms—although this says nothing about spacetime interactions between two or more strings. It is rather easy to see that the string states we have called gravitons really do have to do with small changes in the spacetime metric. In fact, if we consider a small change $h_{\mu\nu}(X)$ which is a plane wave of the form (7.124), then the change in S_{E} is

$$\begin{aligned}\delta S_{\mathrm{E}}(\boldsymbol{k}) &= -\frac{1}{2\pi\alpha'}\int \mathrm{d}z\mathrm{d}\bar{z}\, \epsilon_{\mu\nu}\exp\left[-\mathrm{i}k\cdot X(z,\bar{z})\right]\partial X^{\mu}(z)\bar{\partial}X^{\nu}(\bar{z})\\ &= \frac{1}{4\pi}\int \mathrm{d}z\mathrm{d}\bar{z}\, \mathcal{V}_{\mathrm{g}}(z,\bar{z};\boldsymbol{k}).\end{aligned} \tag{15.238}$$

Apart from the normal ordering, it is given simply by the graviton vertex operator (15.236). The normal ordering was important in making sure that the vertex operator had the desired effect, so it might be as well to fix this up by considering the action to be normal ordered. We have not previously had to consider the action as a quantum operator, but we have dealt at length with the energy–momentum tensor T_{ab} which, according to (15.21), is a derivative of the action. We know from the considerations of §15.3.1 that normal ordering is the right way to make T_{ab} into a well defined operator, so it is at least consistent to suppose that the action should be normal ordered also. In outline, at least, we see that a quantum string interacts with the spacetime metric by emitting and absorbing gravitons. We shall be able to make the outline a little sharper in §15.5.1, which deals with interacting strings.

According to general relativity, $g_{\mu\nu}$ is not an arbitrary metric, but should be a solution of the field equations (4.17). Equations more or less equivalent to these arise in string theory from the requirement of Weyl invariance. The criterion we found in §15.3.3 for the validity of gauge fixing was that both coordinate and Weyl transformations should be valid quantum symmetries. In particular, this required that the energy–momentum tensor retains both of its classical properties $\nabla^{a}T_{ab} = 0$ and $T_{a}^{a} = 0$. When the worldsheet field theory is a non-interacting one, with the spacetime metric $g_{\mu\nu} = \eta_{\mu\nu}$, a sufficient condition for this is that $d = 26$. When the worldsheet field theory is the interacting one specified by (15.237), it turns out that further conditions are necessary. Here, I cannot enter into the technicalities that are needed to investigate this question properly, but some useful insight can be gained from the following considerations. Like the interacting field theories of chapter 9, the action (15.237) is too complicated to allow exact calculations, and we must resort to perturbation theory. In this case, the parameter α' can be treated, at least in a formal way, as a small coupling constant. To see how this works, consider the gauge-fixed action (15.31), written in terms of the coordinates σ and τ, but modified to allow for a general spacetime metric. For a closed string, with σ running from 0 to 2π, it is

$$S = -\frac{1}{4\pi\alpha'}\int \mathrm{d}\tau \int_0^{2\pi} \mathrm{d}\sigma\, g_{\mu\nu}(X)\partial_a X^\mu \partial^a X^\nu . \tag{15.239}$$

Classically, we might consider shrinking the string to a point, so that its spacetime position is the same for all values of σ and $X^\mu(\tau,\sigma) = x^\mu_{\mathrm{cl}}(\tau)$. In that case, the action is identical to (4.2) with $m = 1/\alpha'$. (The difference between this classical mass and the quantum-mechanical mass given by (15.200) is just one of the things we would have to take account of in a more rigorous treatment.) The Euler–Lagrange equation for $x^\mu_{\mathrm{cl}}(\tau)$ is just the geodesic equation (4.4) for the spacetime trajectory of a classical particle.

Suppose, then, that the expectation value $\langle X^\mu(\tau,\sigma)\rangle$ is this classical path $x^\mu_{\mathrm{cl}}(\tau)$, and write

$$X^\mu(\tau,\sigma) = x^\mu_{\mathrm{cl}}(\tau) + \sqrt{\alpha'}Y^\mu(\tau,\sigma). \tag{15.240}$$

By substituting this into the action (15.239) and expanding in powers of α', we get

$$S(X) = S(x_{\mathrm{cl}}) - \frac{1}{4\pi}\int \mathrm{d}\tau \int_0^{2\pi} \mathrm{d}\sigma\, g_{\mu\nu}(x_{\mathrm{cl}})\partial_a Y^\mu \partial^a Y^\nu + \dots . \tag{15.241}$$

The first term, proportional to α'^{-1} is a constant, which is irrelevant to the quantum theory of the fields Y^μ; the term proportional to $\alpha'^{-1/2}$ is zero because $x_{\mathrm{cl}}(\tau)$ is an extremum of S; the next term, which is independent of α', describes a string moving in a spacetime with the classical metric $g_{\mu\nu}(x_{\mathrm{cl}})$. The remaining terms are proportional to positive powers of α', and can be treated by the methods of perturbation theory outlined in chapter 9. To cut a longish story short, the extra condition needed for T^a_a to vanish is of the form

$$\mathcal{R}_{\mu\nu} = \alpha'\mathcal{T}_{\mu\nu} + \mathrm{O}(\alpha'^2) \tag{15.242}$$

where $\mathcal{R}_{\mu\nu}$ is the spacetime Ricci tensor. On the right-hand side, $\mathcal{T}_{\mu\nu}$ and the higher-order contributions can be interpreted in terms of the spacetime stress tensor in (4.20) for 'stringy' matter, together with further geometrical contributions involving the spacetime Riemann tensor $\mathcal{R}_{\mu\nu\sigma\tau}$.

Superficially, we can draw several important conclusions. First, comparing (15.242) with (4.20), we see that the constant α', which determines the string tension, can loosely be identified with the constant $\kappa = 8\pi G$ that we determined in (4.23). In that case, the mass formula (15.79) tells us that the masses of particles corresponding to vibrating states of the string are $M = (2/\sqrt{\alpha'})(N-1)^{1/2} \sim (N-1)^{1/2}M_{\mathrm{Pl}}$, where $M_{\mathrm{Pl}} = (\hbar c/G)^{1/2} = 2.176 \times 10^{-8}$ kg is the *Planck mass*, whose equivalent energy is $E_{\mathrm{Pl}} = M_{\mathrm{Pl}}c^2 = 1.2 \times 10^{19}$ GeV. Particles with such large masses could not be created in the laboratory. We would therefore hope to be able to identify observed particles with the *massless* states of the string, their relatively small masses being generated by symmetry-breaking or higher-order quantum effects of some kind. Second, the expansion parameter has, in natural units, the dimensions of (energy)$^{-2}$ or (length)^2. The overall magnitude of terms involving higher powers of α' will therefore be determined by a dimensionless parameter such as $\alpha' E^2$ or α'/L^2, where E is the characteristic energy or L is the characteristic length of a particular phenomenon that we want to investigate. This dimensionless parameter will be small if E is smaller than the Planck energy, or if L is larger than the Planck length $L_{\mathrm{Pl}} = (G\hbar c^{-3})^{1/2} = 1.615 \times 10^{-35}$ m. Loosely, indeed, we can identify α' as the characteristic 'string scale'. Observed phenomena, whose characteristic energies per particle are much smaller than E_{Pl} and whose characteristic lengths are much greater than L_{Pl} should be describable in terms of an *effective low-energy theory* or an *effective large-distance theory*, which can be derived as an approximation to string theory by treating α' as very small. From the considerations outlined above, general relativity does seem to emerge as the effective large-distance theory of gravity. The standard model of particle physics certainly does not emerge from bosonic string theory as an effective low-energy theory. Whether it can be derived from some more sophisticated string theory is at present an open question.

Although these conclusions are substantially correct, the argument I used to motivate them cannot be taken at face value, for at least two reasons. One is that the metric $g_{\mu\nu}$ applies to a 26-dimensional spacetime. To obtain a theory of gravity in four dimensions, we must appeal to something like the Kaluza-Klein idea of compactification. The four-dimensional gravitational constant G will be given in terms of the 26-dimensional one by a relation analogous to (8.58). Another difficulty can be seen by examining the stress tensor (14.90) for a scalar field. If the particles associated with this field are massless and do not interact, then $V(\phi)$ is zero. If so, then the constant κ in the field equations $\mathcal{R}_{\mu\nu} = \kappa\mathcal{T}_{\mu\nu}$ could be removed entirely by redefining ϕ as $\kappa^{-1/2}\phi$. Therefore, we can deduce the real value of κ only if we have some means of deciding on the 'intrinsic scale' of the field ϕ. Consider, for example, the field strength tensor (8.37) of a non-Abelian gauge field. If we change A_μ into $\kappa^{-1/2}A_\mu$, then the ratio of the linear and quadratic terms is measured by $\kappa^{-1/2}g$ rather than g. In practice, it is quantities such as κ/g^2 which can be related unambiguously to α' and, of course, a coupling constant such as g must be deduced from a theory of interacting strings.

15.5 Further Developments

In the last three sections, I have tried to give substance to the idea that all the fundamental particles we know of might be described as different states of vibration of a single basic object—a relativistic string; to indicate how a quantum-mechanical account of gravity is automatically included in this description; and to expose the technical issues that arise in an initial attempt to make such a theory work. Clearly, the free bosonic string is far from giving us a sensible account of the world as we know it: it makes sense only in a 26-dimensional spacetime and its ground state is a tachyon, which is inconsistent with the rather fundamental notion of causality. These difficulties became apparent only after a fairly lengthy investigation, but other shortcomings should have been apparent from the start. The closed bosonic string has a state that we have interpreted as a graviton and the open string has a massless state that we might try to interpret as a gauge boson, but there are no fermionic degrees of freedom that might provide us with quarks and leptons and no internal degrees of freedom that might correspond to isospin and the like. Lastly, we have no idea of what might cause these particles to interact.

In this final section, I can offer only a superficial glimpse of some of the ideas that have been tried out in the attempt to construct a theory which might qualify as the ultimate 'theory of everything'.

15.5.1 String interactions

I know no simple way of deriving from first principles the formalism that is used in string theory to account for the scattering of strings, so I shall simply set out the essence of the procedure which has been found to work. As foreshadowed by our discussion of vertex operators in §15.4.2, the basic assumption is that an elementary scattering process corresponds to a single worldsheet with enough 'ends' to account for all the incoming and outgoing particles. For the elastic scattering of two closed strings (that is, the process whereby two incoming particles become two outgoing particles), the simplest worldsheet might look like figure 15.4(a) as seen by a spacetime observer. We have seen, though, that the worldsheet geometry, described by the metric γ_{ab}, is a more or less arbitrary matter, because different metrics are related to each other by a combination of diffeomorphisms and Weyl transformations, both of which are gauge symmetries. The cylindrical worldsheet of figure 15.2(a) could be made, by adjusting the metric appropriately, to look like the disc of figure 15.2(b), where one end of the cylinder is an infinitesimal circle at the origin, while the other is a very large circle at infinity. Equally, it is possible to choose the metric so that both ends become infinitesimal circles. In fact, if we identify each end as a single point, then the worldsheet has the topology of a sphere. Similarly, the metric on the worldsheet of figure 15.4(a) can be adjusted to look like figure 15.4(b), where all four ends have become infinitesimal circles. This is especially convenient, because the

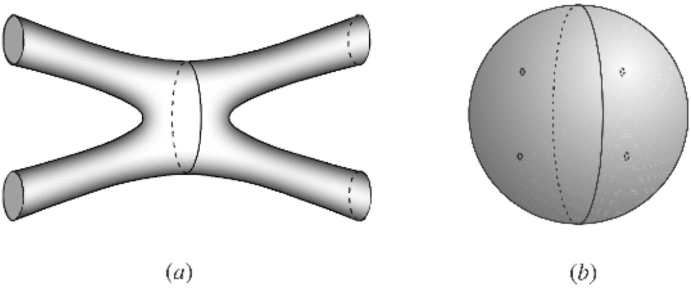


Figure 15.4. A world sheet with four 'ends', which can be interpreted as the two incoming and two outgoing particles in an elastic collision. As embedded in spacetime, the worldsheet might appear as in (*a*), but with a suitable choice of the worldsheet metric, the internal geometry is that depicted in (*b*), where the incoming and outgoing particles are infinitesimal punctures in a spherical surface.

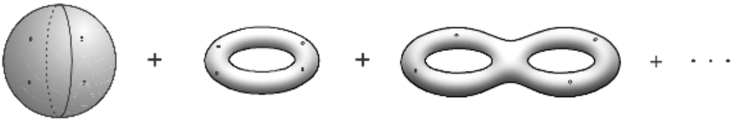


Figure 15.5. The first few worldsheets in an infinite series which gives the total elastic scattering amplitude for two strings. Each one is a compact 2-dimensional surface with four infinitesimal punctures.

vertex operators such as (15.229) and (15.236) which we have available to us to represent the states of the incoming and outgoing particles refer to just one point of the worldsheet, rather than to a circle of finite size.

Compared with the Feynman diagrams of figure 9.3, the worldsheet of figure 15.4 is the analogue of just the first diagram. The sum of diagrams which represents the complete scattering amplitude can be envisaged as in figure 15.5. There is just one diagram for each possible topology of the worldsheet. Like the Feynman diagrams, these pictures of worldsheets merely give a visual impression of the formula that is used to calculate the scattering amplitude. The formula is this:

$$
\begin{aligned}
&\mathcal{S}_{i_1\cdots i_n}(k_i, \ldots, k_n) \\
&\qquad = \mathcal{N} \sum_{\substack{\text{worldsheet}\\ \text{topologies}}} \int \mathcal{D}X\, \mathcal{D}\gamma\, \exp\left[-S_{\mathrm{E}}(X, \gamma) - \lambda\chi\right] V_{i_1}(k_1) \cdots V_{i_n}(k_n).
\end{aligned}
\tag{15.243}
$$

It is the string-theory analogue of the second expression on the right-hand side of (9.13), generalized to include a total of *n* incoming and outgoing particles. In particular, the differential operators $(\Box + m^2)$ serve to cancel out the external propagators in diagrams such as those of figure 9.3, and this corresponds loosely to the fact that the cylindrical 'legs' of figure 15.4(a) can be contracted to the points in figure 15.4(b). The vertex operators in (15.243) are integrated over all positions on the worldsheet:

$$V_i(k) = \int \mathrm{d}^2z\, \gamma^{1/2}(z, \bar{z})\mathcal{V}_i(z, \bar{z}; k) \tag{15.244}$$

where the factor $\gamma^{1/2}$ allows for the fact that we may not be able to choose a flat worldsheet metric when fixing the gauge. Evidently, the factors $\mathrm{e}^{\pm \mathrm{i}k\cdot x}$ in (9.13) are reflected by the exponentials in the vertex operators. The values of k_0 in the vertex operators are given by $k^0 = \pm\left(|\boldsymbol{k}|^2 + M_i^2\right)^{1/2}$, where M_i^2 is the mass^2 for species i, as obtained from the appropriate mass formula; the positive sign corresponds to an incoming particle and the negative sign to an outgoing particle.

The quantity χ, which appears in (15.243) multiplied by a constant λ, is the *Euler characteristic* of the worldsheet, given by

$$\chi = \frac{1}{4\pi}\int \mathrm{d}^2\sigma\, \gamma^{1/2}R(\sigma) \tag{15.245}$$

where σ^i are any convenient coordinates on the Euclidean worldsheet (on which γ is positive) and R is the 2-dimensional Ricci scalar. As indicated in exercise 15.1, this is the 2-dimensional equivalent of the Einstein–Hilbert action (4.16), but $\gamma^{1/2}R$ is a total divergence and makes no contribution to the equations of motion. On a compact surface such as a sphere or a torus, it is not possible to have a metric that is flat everywhere, so in general χ does not vanish. On the other hand, its value depends only on the topology of the worldsheet and is independent of the metric. For a sphere, it is equal to 2 (see exercise 15.12) and for the tori shown in figure 15.5, it is $\chi = 2 - 2G$, where the *genus* G is the number of holes. More generally, there are several other topologies to be accounted for and a more general formula for χ, but χ is always an integer. We see that each hole added to the worldsheet gives rise to a factor $g^2 = \mathrm{e}^{2\lambda}$ in the scattering amplitude (15.243), so g is the analogue of a coupling constant associated with a vertex in a field-theoretic Feynman diagram and the perturbation series represented by the sum of string diagrams is, in a sense, an expansion in powers of g.

Consider now the effect on a scattering amplitude (15.243) of a small change in the spacetime metric $g_{\mu\nu}$, which corresponds to a small change δS_{E} in the action. The small change in the scattering amplitude is

$$\begin{aligned}&\delta\mathcal{S}_{i_1\cdots i_n}(k_i, \ldots, k_n)\\ &\quad = -\mathcal{N}\sum_{\substack{\text{worldsheet}\\ \text{topologies}}}\int \mathcal{D}X\,\mathcal{D}\gamma\, \exp[-S_{\mathrm{E}}(X,\gamma) - \lambda\chi]\,\delta S_{\mathrm{E}}V_{i_1}(k_1)\cdots V_{i_n}(k_n).\end{aligned} \tag{15.246}$$

In general, δS_{E} might be a linear superposition of the Fourier modes $\delta S_{\mathrm{E}}(\boldsymbol{k})$ shown in (15.238), which is just a graviton vertex operator. This vertex operator for a graviton appears in (15.246) in just the same way as the vertex operators $V_{i_1}(k_1)\cdots V_{i_n}(k_n)$ for the other particles, so we see a little more clearly that a small change in $g_{\mu\nu}$ is equivalent to the emission or absorption of a graviton. A large change in $g_{\mu\nu}$ can be built up from many small ones and is equivalent to a coherent superposition of many gravitons. In this sense, we can say that to change the metric in (15.237) is not to change the theory, but to study a different state of the same theory—a state with a different number of gravitons.

The same argument must apply to the other massless particles. That is to say, the emission and absorption of particles other than gravitons should be equivalent to changing other 'backgrounds' in the action analogous to $g_{\mu\nu}(X)$. Consider, in particular, adding to S_{E} a term of the form

$$\Delta S_{\mathrm{E}} = \int \mathrm{d}^2\sigma\, \gamma^{1/2} R\Phi(X). \tag{15.247}$$

Using the Euclidean version of the result of exercise 15.1, namely $\gamma^{1/2}R = \partial^a\partial_a\Omega$ and an integration by parts, we can write

$$\Delta S_{\mathrm{E}} = \int \mathrm{d}^2\sigma\, \Omega\partial^a\partial_a\Phi(X) \tag{15.248}$$

and a small change in $\Phi(X)$ proportional to $\mathrm{e}^{-\mathrm{i}k\cdot X}$ produces a small change in the action of

$$\delta\Delta S_{\mathrm{E}} = -\int \mathrm{d}^2\sigma\, \Omega k_\mu k_\nu \partial^a X^\mu \partial_a X^\nu \mathrm{e}^{-\mathrm{i}k\cdot X} \tag{15.249}$$

which is a linear combination of vertex operators. Changing $\Phi(X)$ by a constant is equivalent to changing the constant λ, or the coupling constant $g = \mathrm{e}^\lambda$. Therefore, different values of g also correspond not to different theories, but to different states of the same theory. The conclusion is that string theory has, in fact, no adjustable constants (although it has many possible states, which in practice might amount to much the same thing). Earlier on, I showed that the constant α' is related to the Planck mass, albeit indirectly when we take account of some mechanism such as compactification to reduce the number of observable dimensions to four. However, the actual value of α' is not physically meaningful, for the following reason. Suppose that string theory does indeed describe our world, and that its physical implications could be worked out in detail. We ought then to be able to calculate, say, the mass of the proton, whose equivalent energy is about 1 GeV. Since α' is the only dimensionful parameter, we would get an answer of the form $m_{\mathrm{p}} = \mathcal{M}_{\mathrm{p}}\alpha'$, where $\mathcal{M}_{\mathrm{p}}$ is a dimensionless number. To say that α' is of the order of the Planck mass $M_{\mathrm{Pl}} \simeq 10^{19}$ GeV is to say that the number $\mathcal{M}_{\mathrm{p}}$ is of the order of 10^{-19}. In fact, the value of any physical quantity can be determined only as a multiple of some standard quantity such as m_{p}, so only dimensionless ratios such as $\mathcal{M}_{\mathrm{p}}$ have physical meaning. If it is correct, then, string theory is a theory within which all measurable physical quantities can in principle be calculated with no adjustable parameters. This is a large part of its attraction as a candidate for the 'theory of everything'.

15.5.2 Superstrings

One way of endowing the string with internal degrees of freedom is to enlarge the two-dimensional field theory living on the worldsheet by adding more fields. If the fields are fermionic, then we might hope to find amongst the states of the string particles which behave as fermions in spacetime, although the connection between these two ideas is not entirely straightforward. An idea which has proved particularly fruitful is that the bosonic degrees of freedom, the spacetime coordinates X^{μ}, and the fermionic ones should be related by a supersymmetry, such as we discussed in §12.7. Indeed, it is in this string theory context that supersymmetry was first discovered. On a curved worldsheet, the action which generalizes (15.15) is

$$S = -\frac{1}{4\pi}\int_{-\infty}^{\infty} \mathrm{d}\tau \int_{0}^{\ell} \mathrm{d}\sigma \, (-\gamma)^{1/2}\left[\frac{1}{\alpha'}\gamma^{ab}\partial_a X_{\mu}\partial_b X^{\mu} - \mathrm{i}\bar{\Psi}_{\mu}\rho^a\nabla_a\Psi^{\mu}\right] \tag{15.250}$$

the last term being a 2-dimensional version of the generally covariant action (7.147) with $m = 0$. The new fermionic fields $\Psi^{\mu}(\tau, \sigma)$ are a set of *d Majorana* spinors, and this action is rather like a set of *d* copies of the supersymmetric Wess–Zumino model (12.83). The new theory defined in this way has a larger gauge symmetry than the bosonic string, consisting of diffeomorphism invariance, Weyl invariance and a local supersymmetry. (To be accurate, the *local* supersymmetry holds if we extend the action by adding a 2-dimensional 'gravitino', which disappears again upon fixing the gauge.) This enlarged gauge symmetry can be fixed in much the same way that we studied for the bosonic string; its remnant on the flat worldsheet is an enlarged version of conformal symmetry, called a *superconformal symmetry*. The condition for the classical gauge symmetry to remain valid as a quantum symmetry leads, as it turns out, to a critical spacetime dimension $d = 10$.

On the flat worldsheet, the 2-dimensional Dirac matrices ρ^a, with the anticommutation relations $\{\rho^a, \rho^b\} = 2\eta^{ab}$ and the charge conjugation matrix, with the property $C\rho^{a\mathrm{T}}C^{-1} = -\rho^a$ can be chosen as

$$\rho^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \rho^1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \rho^1. \tag{15.251}$$

If we write the two components of Ψ^{μ} as $\widetilde{\psi}^{\mu}$ and $\mathrm{i}\psi^{\mu}$, then the Majorana condition $\Psi^{\mu\mathrm{c}} \equiv C\rho^0\Psi^{\mu *} = \Psi$ (see §7.5) becomes

$$\Psi^{\mu\mathrm{c}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \widetilde{\psi}^{\mu} \\ \mathrm{i}\psi^{\mu} \end{pmatrix}^{*} = \begin{pmatrix} \widetilde{\psi}^{\mu *} \\ \mathrm{i}\psi^{\mu *} \end{pmatrix} = \Psi = \begin{pmatrix} \widetilde{\psi}^{\mu} \\ \mathrm{i}\psi^{\mu} \end{pmatrix} \tag{15.252}$$

so the components ψ^μ and $\widetilde{\psi}^\mu$ are real. The 2-dimensional Dirac equation $\partial\!\!\!/\Psi = 0$ can easily be found to imply

$$\bar{\partial}\psi^\mu \equiv \tfrac{1}{2}(\partial_\sigma + \partial_\tau)\,\psi^\mu = 0 \qquad \partial\widetilde{\psi}^\mu \equiv \tfrac{1}{2}(\partial_\sigma - \partial_\tau)\,\widetilde{\psi}^\mu = 0 \tag{15.253}$$

so ψ^μ, like ∂X^μ, is a real, right-moving field, while $\widetilde{\psi}^\mu$, like $\bar{\partial}X^\mu$, is a real, left-moving field. In this way, we see that the numbers of bosonic and fermionic degrees of freedom match up in a way that makes a *worldsheet supersymmetry* possible. This is, however, by no means enough to guarantee the existence of a *spacetime supersymmetry*, which would mean that the physical states of the string fall into supersymmetry multiplets analogous to those we discussed in §12.7.

On a flat, Minkowskian worldsheet, the action can now be written as

$$S = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\tau \int_0^\ell \mathrm{d}\sigma \left[\frac{2}{\alpha'}\,\partial X_\mu \bar{\partial} X^\mu + \mathrm{i}\psi_\mu \bar{\partial}\psi^\mu - \mathrm{i}\widetilde{\psi}_\mu \partial \widetilde{\psi}^\mu\right]. \tag{15.254}$$

Each bosonic field $X^\mu(\tau,\sigma)$ is to be identified as a spacetime coordinate of the point (τ,σ) of the worldsheet, and must therefore have a unique value at each point. However, the internal degrees of freedom ψ^μ and $\widetilde{\psi}^\mu$ need not be single-valued. All we require is that the Lagrangian density have a unique value. For a closed string, this means that ψ^μ may be either periodic or antiperiodic:

$$\psi^\mu(\tau, \sigma + \ell) = \pm\psi^\mu(\tau,\sigma) \tag{15.255}$$

and similarly for $\widetilde{\psi}^\mu$. For an open string, the boundary term analogous to (15.17) involved in the derivation of the Dirac equation can be made to vanish by imposing the conditions

$$\psi^\mu(\tau,0) = \widetilde{\psi}^\mu(\tau,0) \qquad \psi^\mu(\tau,\ell) = \pm\widetilde{\psi}^\mu(\tau,\ell). \tag{15.256}$$

Fields which satisfy (15.255) or (15.256) with the + sign are said to have *Ramond* (R) boundary conditions (after P Ramond); with the − sign, they are said to have *Neveu-Schwarz* (NS) boundary conditions (after A Neveu and J H Schwarz).

Because of these different boundary conditions, the full Hilbert space $\mathcal{H}$ of the superstring contains several topological sectors, akin to those we met in (13.1) for soliton-bearing field theories. Each sector can be further subdivided according to the value of what is known (for historical reasons that need not concern us) as *G-parity*. This G-parity is even or odd, according to the number of fermionic creation operators that are needed in an equation such as (15.212) to create a given state. The question arises whether all of these sectors can be represented in the physical Hilbert space $\mathcal{H}_{\text{phys}}$. When interactions are allowed, it turns out that they cannot: there are consistency requirements which allow only certain combinations of sectors to appear in $\mathcal{H}_{\text{phys}}$. To say exactly what these requirements are needs some technology that I do not have the space to develop in detail, but they amount to demanding that an expression analogous to (15.243) should give unambiguous and fully gauge-invariant results for superstring scattering amplitudes. The

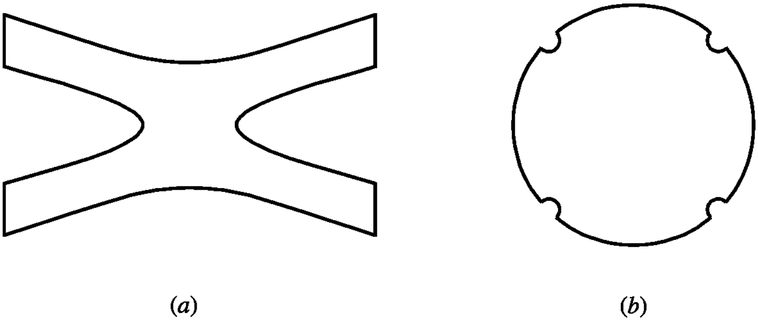


Figure 15.6. The open-string version of figure 15.4. As viewed in spacetime, the elastic scattering of two strings might look as in (a), but with a suitable choice of the worldsheet metric the internal geometry is that shown in (b), where the 'ends' of the string corresponding to incoming and outgoing particles are infinitesimal semicircles set into the boundary of a disc.

imposition of these requirements is called (after F Gliozzi, J Scherk and D Olive) the *GSO projection*. In the case of a theory constructed only from closed strings, the net result is that there are exactly two fully consistent theories, called the *type IIA* and *type IIB* theories. Both theories have the feature that, although tachyonic states appear in the full Hilbert space, they are excluded from the physical Hilbert space. This is clearly a great advantage. It is also true of both theories that the physical states form multiplets of a spacetime supersymmetry, and this too is perceived by practitioners as an attractive feature. The two theories differ in respect of the way in which sectors associated with the left-moving modes are combined with those associated with the right-moving modes, and this in turn affects the grouping of their particle states into supermultiplets. In particular, the type IIA theory is *non-chiral*. Roughly, this means that there is a symmetry between states of positive and negative helicity. The standard electroweak model (§12.2), for example, does not have this 'left-right' symmetry because the weak isospin doublets contain only left-handed particles, and is said to be a *chiral* theory. The spectrum of massless particle states of the type IIB superstring is chiral in a similar sense.

For an open string, it proves possible to introduce isospin-like degrees of freedom by assigning integer labels, say $i, j = 1 \dots n$ to the ends of the string. They are called *Chan–Paton* degrees of freedom (after H M Chan and J E Paton). This apparently trivial device has no effect on the worldsheet field theory for a free string, but it does affect the number of distinct worldsheets that appear in the formula (15.243) for scattering amplitudes in the interacting theory. Let us denote a state of an open string with Chan–Paton indices i and j by $|\Psi; i, j\rangle$, where Ψ represents all the other degrees of freedom we know about. We can restrict the

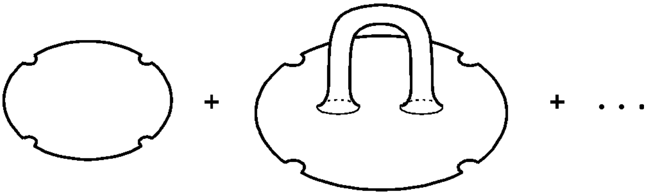


Figure 15.7. A correction to the open-string scattering amplitude. The new worldsheet is a disc with a handle attached, which can be interpreted as a virtual closed string.

allowed states to be linear superpositions of the form

$$|\Psi; a\rangle = \sum_{i,j} |\Psi; i, j\rangle T^a_{ij} \tag{15.257}$$

where the $n \times n$ matrices T^a are the generators of a Lie group, such as we met in §8.2. These otherwise indistinguishable particle states transform into one another under the action of the Lie group, which thereby becomes an internal symmetry of the string theory. The worldsheets of interacting open strings have boundaries corresponding to the long edges of the single string of figure 15.1(a). With a suitable choice of the worldsheet metric, the simplest ones look like discs, with infinitesimal semicircles arranged round their edges, which are the ends of the incoming and outgoing strings, as sketched in figure 15.6. As indicated in figure 15.7, higher-order worldsheets include those with 'handles' attached to the disc, and we see that one of these handles is a virtual closed string. The fact that there are virtual processes involving closed strings suggests that a theory which allows open strings as the real incoming and outgoing particles ought also to allow for real closed-string particles. Indeed, it can be shown that closed strings *must* also be included if the requirement of unitarity is to be satisfied (that is, if the total probability of observing *some* final state is to be 1). A theory of this kind is a *type I* superstring. It would be helpful if the calculated scattering amplitudes turned out to be finite. For reasons related to the gauge anomalies that I mentioned in chapter 9, this is found to be true only for one special choice of the Lie group associated with the Chan-Paton degrees of freedom, namely SO(32), which is the group of rotations in 32 dimensions.

At this point, we have three apparently different versions of the superstring, which are thought to be mathematically sound. All of them are free of tachyons and all of them exhibit spacetime supersymmetry. There are two further versions known which share these properties, called *heterotic* strings. These are closed strings, and on a flat worldsheet, they have an action of the form

$$S = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\tau \int_0^{\ell} \mathrm{d}\sigma \left[\frac{2}{\alpha'}\partial X_\mu \bar{\partial} X^\mu + \mathrm{i}\psi_\mu \bar{\partial}\psi^\mu - \mathrm{i}\sum_{i=1}^{n} \widetilde{\lambda}^i \partial \widetilde{\lambda}^i\right]. \tag{15.258}$$

This action differs from (15.254) in that the left-moving fermions $\widetilde{\lambda}^i$ do not form a spacetime vector, but are simply a collection of n fields. As with the bosonic string, the left- and right-moving modes are independent of each other. The right-moving bosons and fermions of the heterotic string are still related by a worldsheet supersymmetry, but the left-moving ones are not. Thus, the gauge symmetry is smaller. Because the right-moving part of the theory is identical to the previous superstrings, the condition for its superconformal central charge to vanish is still $d = 10$. The number n of left moving fermions is determined by the condition that the central charge for the left-moving modes should also vanish. It might seem that n should be 10, but this is not so, because the smaller gauge symmetry results in fewer ghosts when the gauge is fixed. The actual number turns out to be $n = 32$. The heterotic string has an internal symmetry, consisting of rearrangements of the 32 fermionic fields λ^i. In fact, if we regard these fields as the components of a vector $\boldsymbol{\lambda}$ in a 32-dimensional Euclidean space, then the dot product $\boldsymbol{\lambda} \cdot \widetilde{\partial}\boldsymbol{\lambda}$ which appears in the action is invariant under rotations in this space, so the internal symmetry group is again SO(32). This is true, at least, if all the λ^i are completely equivalent. They will not be completely equivalent if they have different boundary conditions, however. Two possibilities are found to lead to mathematically consistent theories. One is that all the λ^i have the same boundary conditions, in which case the symmetry group is indeed SO(32). The other possibility is to assign boundary conditions independently to two groups of 16 fields each. The symmetry group of this type of heterotic string goes by the name of $\mathrm{E}_8 \times \mathrm{E}_8$.

All in all, there are five known superstring theories which seem to be mathematically sound. It might seem that these would constitute five competing candidates for the theory of everything, but recent investigations point to a more intriguing possibility, as I shall shortly try to explain.

15.5.3 The ramifications of compactification

Superstrings are mathematically well defined in ten, rather than twenty-six spacetime dimensions, but this is still too many—to the tune of six. Somehow, we must explain why only four are apparent to us, and the Kaluza-Klein idea that we touched on in §8.5 provides a starting point. In general terms, the 10-dimensional spacetime manifold must be split into a product $L^4 \times C^6$, which means that each point of the 4-dimensional manifold L^4, whose dimensions are large, is really a 6-dimensional manifold C^6 whose dimensions are compactified to a small size. One question that arises immediately is, what sort of manifold is C^6? The possibilities are, as it were, manifold. In figure 8.1, where only one dimension is compactified, C^1 is a circle. If two dimensions are compactified, the simplest possibility to deal with is that C^2 is a torus: we simply have to say that every function on the torus is periodic in both compact dimensions. This idea is straightforward to generalize to a 6-dimensional torus C^6, but this is by no means the only possibility. Amongst other manifolds, those known as orbifolds,

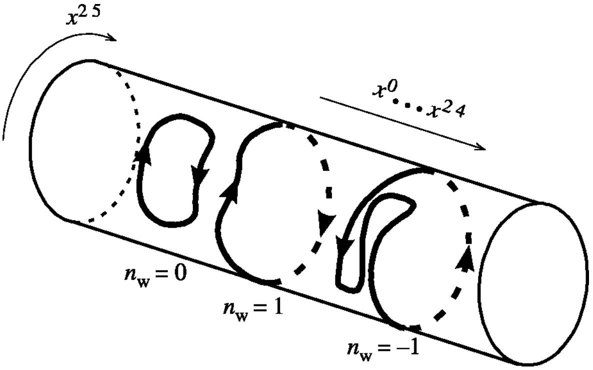


Figure 15.8. Schematic illustration of closed-string configurations in a spacetime with one compacted dimension. The string may wind n_W times around the compact dimension, and possible configurations with $n_W = 0$, 1 and -1 are shown.

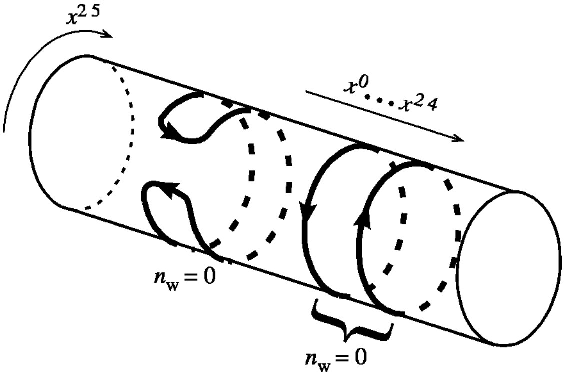


Figure 15.9. A closed string with a winding number of 0 may intersect itself and subsequently split to form two strings with opposite winding numbers.

orientifolds and Calabi-Yau manifolds seem to have advantageous properties, but I cannot enter into them here. To illustrate the sorts of considerations that arise, I shall look at what happens when just one of the dimensions of the 26-dimensional spacetime of the bosonic string is compactified to a circle.

Say that the compactified dimension is X^{25} and that the circumference of the circle is $2\pi r$. A closed string might wind n_W times around the 'cylinder', where n_W is any positive or negative integer or zero. Figure 15.8 shows closed strings with winding numbers of 0, 1 and -1. Were we content to deal only with noninteracting strings, then it might be possible to discard all the possibilities except $n_W = 0$. However, if we allow strings to interact by joining and splitting as in figure 15.4(a), then all possibilities must be included. For example, figure 15.9

shows that a string of winding number 0 might decay into two strings with winding numbers 1 and -1. All of these possibilities must be allowed for in the Hilbert space of a free string, and there are two ways in which our earlier considerations are modified. First, the total change in the coordinate X^{25} as σ varies from 0 to 2π is

$$\int_0^{2\pi} \mathrm{d}\sigma\, \partial_\sigma X^{25}(\tau,\sigma) = X^{25}(\tau,2\pi) - X^{25}(\tau,0) = 2\pi r n_\mathrm{w}. \tag{15.259}$$

Second, the eigenvalues k^{25} of the spacetime momentum p^{25} have the discrete values

$$k^{25} = n_k/r \tag{15.260}$$

just as the momentum of an ordinary non-relativistic particle confined to a finite-sized box is quantized. We see this directly by observing that if the whole string is moved a distance $2\pi r$ in the X^{25} direction without changing its shape, then its state is left unchanged. Therefore, vertex operators such as (15.229) and (15.236) must be left unchanged if we replace X^{25} by $X^{25} + 2\pi r$.

These two facts can be accommodated in our formalism by writing X^{25} as the sum of a right-moving part X_R and a left-moving part X_L:

$$X_\mathrm{R}(\sigma-\tau) = x_\mathrm{R} + \tfrac{1}{2}\alpha' p_\mathrm{R}(\tau-\sigma) + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty} \frac{1}{n}\alpha_n^{25} \mathrm{e}^{-\mathrm{i}n(\tau-\sigma)} \tag{15.261}$$

$$X_\mathrm{L}(\sigma+\tau) = x_\mathrm{L} + \tfrac{1}{2}\alpha' p_\mathrm{L}(\tau+\sigma) + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty} \frac{1}{n}\widetilde{\alpha}_n^{25} \mathrm{e}^{-\mathrm{i}n(\tau+\sigma)}. \tag{15.262}$$

The centre-of-mass operator is $x^{25} = x_\mathrm{R} + x_\mathrm{L}$ and the component of spacetime momentum in the X^{25} direction is

$$p^{25} = \tfrac{1}{2}\left(p_\mathrm{R} + p_\mathrm{L}\right). \tag{15.263}$$

When this operator acts on a basis vector with $k^{25} = n_k/r$, we can write

$$\tfrac{1}{2}\left(p_\mathrm{R} + p_\mathrm{L}\right) = \frac{1}{r} n_k. \tag{15.264}$$

The two operators p_R and p_L are not independent, because (15.259) shows that

$$\tfrac{1}{2}\left(p_\mathrm{L} - p_\mathrm{R}\right) = (r/\alpha')n_\mathrm{w}. \tag{15.265}$$

Obviously, they are not equal unless $n_\mathrm{w} = 0$, and the expansion coefficients α_0^{25} and $\widetilde{\alpha}_0^{25}$ must be identified as

$$\alpha_0^{25} = (\alpha'/2)^{1/2} p_\mathrm{R} \qquad \widetilde{\alpha}_0^{25} = (\alpha'/2)^{1/2} p_\mathrm{L} \tag{15.266}$$

in place of (15.69). As a result, the expression (15.198) for the Virasoro generator L_0 and the analogous expression for $\widetilde{L}_0$ become

$$L_0 = -\frac{\alpha'}{4}\left(M^2 - p_{\mathrm{R}}^2\right) + N - 1 \qquad \widetilde{L}_0 = -\frac{\alpha'}{4}\left(M^2 - p_{\mathrm{L}}^2\right) + \widetilde{N} - 1 \quad (15.267)$$

where $M^2 = (p^0)^2 - \sum_{\mu=1}^{24}(p^\mu)^2$ is the mass^2 of a particle in the 25 non-compactified dimensions. The levels N and $\widetilde{N}$ are still given by the two expressions (15.202) but they are now not equal. In fact, for a state of definite momentum $p^\mu = k^\mu$, the two constraints $L_0|\psi\rangle = \widetilde{L}_0|\psi\rangle = 0$ now imply

$$M^2 = \frac{1}{r^2}n_k^2 + \frac{r^2}{\alpha'^2}n_{\mathrm{w}}^2 + \frac{2}{\alpha'}\left(N + \widetilde{N} - 2\right) \qquad (15.268)$$

$$N - \widetilde{N} = n_k n_{\mathrm{w}} \qquad (15.269)$$

as some simple arithmetic using (15.264) and (15.265) will show.

As we discussed in §15.4.3, the most interesting particle states are those for which $M^2 = 0$. The expression (15.268) for M^2 can vanish only if the first two terms add to form an integer times $(2/\alpha')$ and for general values of r the only possibility is $n_k = n_{\mathrm{w}} = 0$. In this case, the massless particle states are the same as those we found for a non-compactified spacetime. However, a polarization tensor such as (15.218) for the graviton has a different interpretation, because the values $\mu, \nu = 25$ do not refer to observable spacetime directions. In the 25-dimensional spacetime, the components $\epsilon_{\mathrm{g}}^{\mu\nu}$ give the polarizations of a graviton, while $\epsilon_{\mathrm{g}}^{\mu\,25}$ (which is equal to $\epsilon_{\mathrm{g}}^{25\,\mu}$) corresponds to a vector particle, or 'photon'. We see, from a slightly different point of view, the same phenomenon that we met in (8.57). The 26-dimensional graviton becomes a 25-dimensional graviton plus a 25-dimensional gauge field; the remaining component $\epsilon^{25\,25}$ is the string-theory analogue of the constant g_{55}.

New phenomena, specific to strings, start to become apparent when we observe that there is a special value of r, namely $r = \sqrt{\alpha'}$, for which the mass formula can be written as

$$M^2 = \frac{1}{\alpha'}\left[(n_k - n_{\mathrm{w}})^2 + 4N - 4\right]. \qquad (15.270)$$

There are now several more values of n_k and n_{w} for which this vanishes. The string spectrum contains more massless states which, in the dimensionally reduced theory can be interpreted as the gauge bosons of a larger gauge symmetry. More far-reaching is the observation that both (15.268) and (15.269) are unchanged if we interchange the integers n_k and n_{w} and at the same time replace r with α'/r. That is to say, the particle masses resulting from a compactification radius r are exactly the same as those resulting from a compactification radius $\hat{r} = \alpha'/r$. It can be shown that the scattering amplitudes are also exactly the same, so the physical content of the theories obtained using the radii r and $\hat{r}$ seems to be equivalent.

Now, everything that we can actually calculate has to do with the two-dimensional worldsheet field theory. We started from the idea that the fields X^{μ} were the spacetime coordinates of a moving string, but it is legitimate, and is now becoming desirable, to take a different point of view. That is to say, given a working two-dimensional field theory, we can look for an interpretation of this theory in terms of particles propagating through spacetime. The fact that we have used the notation X^{μ} for some of the fields in our theory need not commit us to interpreting precisely these fields as the spacetime coordinates. With this in mind, consider the change of notation

$$\hat{p}_{\mathrm{R}} = -p_{\mathrm{R}} \qquad \hat{x}_{\mathrm{R}} = -x_{\mathrm{R}} \qquad \hat{\alpha}_n^{25} = -\alpha_n^{25} \qquad \hat{n}_k = n_{\mathrm{w}} \qquad \hat{n}_{\mathrm{w}} = n_k. \tag{15.271}$$

All the equations of our two-dimensional quantum field theory are exactly the same when written in terms of the new variables. For example, the commutation relations for the $\hat{\alpha}_n^{25}$ are the same as (15.99) and the definition (15.73) of the Virasoro generators gives $\hat{L}_n = L_n$, because two minus signs cancel in each case. In the compactified theory, equations (15.264) and (15.265) are interchanged, provided that we take $\hat{r} = \alpha'/r$. At the level of the two-dimensional field theory, then, the compactification radii r and $\hat{r}$ are entirely equivalent. However, to get a spacetime interpretation of the 'hatted' theory, we must take the new field

$$\hat{X}^{25}(\tau, \sigma) = \hat{X}_{\mathrm{R}}(\sigma - \tau) + X_{\mathrm{L}}(\sigma + \tau) = -X_{\mathrm{R}}(\sigma - \tau) + X_{\mathrm{L}}(\sigma + \tau) \tag{15.272}$$

to be the one that represents the 25th spacetime coordinate. The situation is reminiscent of one we encountered in chapter 13, where the change of variables (13.78) relates the usual theory of electromagnetism to a *dual* theory in which the interpretations of electric and magnetic fields are interchanged. Here, the transformation specified by (15.271) or (15.272) is called *T-duality*.

Applied to an open string, T-duality has startling consequences, arising from the superficially innocuous boundary condition $\partial_\sigma X^{\mu}(\tau, \sigma) = 0$ at the end-points $\sigma = 0$ and $\sigma = \pi$ (which we chose as a convenient range for an open string). For $\mu = 25$, this says that at the end-points

$$\partial_\sigma X_{\mathrm{R}}(\sigma - \tau) + \partial_\sigma X_{\mathrm{L}}(\sigma + \tau) = 0. \tag{15.273}$$

In the dual description, we find

$$\begin{aligned} \partial_\tau \hat{X}^{25}(\tau, \sigma) &= -\partial_\tau X_{\mathrm{R}}(\sigma - \tau) + \partial_\tau X_{\mathrm{L}}(\sigma + \tau) \\ &= \partial_\sigma X_{\mathrm{R}}(\sigma - \tau) + \partial_\sigma X_{\mathrm{L}}(\sigma + \tau) \\ &= 0 \end{aligned} \tag{15.274}$$

at the end-points, which are still $\sigma = 0$ and $\sigma = \pi$. The Neumann boundary conditions have been replaced by *Dirichlet* boundary conditions. They say that the ends of the string have *fixed* values of $\hat{X}^{25}$: they cannot move in the $\hat{X}^{25}$ direction. Moreover, the total change in $\hat{X}^{25}$ as σ varies from 0 to π is a multiple of $2\pi\hat{r}$, as we can see by relating it to the momentum $p^{25} = n_k/r$ in the 'unhatted' description:

$$\begin{aligned}\int_0^\pi d\sigma\, \partial_\sigma \hat{X}^{25}(\tau,\sigma) &= \int_0^\pi d\sigma\, \partial_\sigma\, [-X_{\mathrm{R}}(\sigma-\tau) + X_{\mathrm{L}}(\sigma+\tau)] \\ &= \int_0^\pi d\sigma\, \partial_\tau\, [X_{\mathrm{R}}(\sigma-\tau) + X_{\mathrm{L}}(\sigma+\tau)] \\ &= \int_0^\pi d\sigma\, \partial_\tau X^{25}(\tau,\sigma) \\ &= 2\pi\alpha' p^{25} \\ &= 2\pi\hat{r}\, n_k. \end{aligned} \tag{15.275}$$

As far as spacetime geometry is concerned, two values of $\hat{X}^{25}$ which differ by a multiple of $2\pi\hat{r}$ are exactly the same. This means that the two ends of the open string must lie on the 25-dimensional hyperplane that corresponds to some fixed value of $\hat{X}^{25}$.

The considerations that have led us to this point apply to the noninteracting worldsheet field theory, for which (15.261) and (15.262) are exact solutions of the equations of motion. More generally, if 'backgrounds' such as the spacetime metric in (15.237) or $\Phi(X)$ in (15.247) are included, one finds that the 25-dimensional hypersurface on which the ends of the open string lie is not flat, but has a shape which depends on the backgrounds. We saw in §15.5.1 that changing these backgrounds is equivalent to the emission and absorption of particles, so this hypersurface is a physical object, which can interact with the strings. It can be interpreted as the 'worldvolume' traced out by the motion of a 24-dimensional membrane. Many such objects, of various dimensions, have been identified by string theorists in recent years—a circumstance offering unrivalled scope to high-energy physicists' propensity for punning terminology ('p-brane', 'brane scan', '... on the brane', ...). The one we have identified here is called a *D24-brane*, the 'D' indicating its relationship to open strings with Dirichlet boundary conditions, but not all of them arise in the same way. The situation is again analogous to one that we met in chapter 13, where we saw that field theories of point particles may also contain soliton-like objects which, from the point of view of perturbation theory, seem very different from the particles. Indeed, the analogy seems to be a strong one. From string theory, one can derive low-energy effective field theories which describe the massless states of strings as point particles; the Einstein field equations (15.242) provide one example. In some cases, soliton solutions to these field theories can be identified with the string-theory branes. We saw, moreover, that apparently different field theories may be related to each other by duality transformations, and that it may be possible to identify the point particles of one theory with the solitons of a dual theory. The same seems to be true of superstring theories. In fact, the current view is that all five of the superstring theories that I mentioned in §15.5.2 are related by various dualities, of which T-duality is one. Further, a small coupling constant, which makes perturbation theory feasible in one field theory, may correspond to a large coupling constant in a dual theory. In string theory, the indications are that the five superstring theories which are known perturbatively constitute different weak-coupling approximations to a single overarching theory, which has been christened *M-theory*. Exactly what the fundamental principles are that would serve to define M-theory in a precise way, independent of perturbative approximations, is not clear, but it does seem that this theory would naturally exist in eleven dimensions, a compactification of the eleventh dimension being one ingredient of the various weak-coupling limits.

Finally, there exists the intriguing possibility that not all of the extra six (or seven) dimensions need be small. It could be that we do not perceive these dimensions because the particles of which the matter familiar to us is made are constrained to live on a three-dimensional 'brane-world'. If so, it has been speculated that particles able to propagate off the brane-world might be created with energies accessible in the laboratory, and that the extra dimensions might be detected in this way. Gravitons able to propagate off the brane-world might cause a detectable correction to the $1/r^2$ law of gravitational attraction at distances perhaps not much smaller than 1mm.

15.6 The Last Word?

Long as this chapter has been, it affords only an elementary glimpse of the theoretical edifice which, in the eyes of its devotees, offers our best current hope of a truly unified theory of the world at its most fundamental level. At the time of writing, this theory is entering the fourth decade of its evolution, but the true nature of the mathematical structure that might eventually emerge can at best be dimly perceived, even by the experts (of whom I am by no means one). The point of view underlying this chapter has been that the truly fundamental objects in nature are one-dimensional strings, but the picture that seems to be emerging calls this view into question. It is quite likely that branes of many different dimensions appear on much the same footing, and that none of these objects can be regarded as truly fundamental. From the point of view of the 11-dimensional M-theory, for example, the type-IIA string is to be interpreted as a 2-dimensional membrane, one of whose dimensions winds around a small, compact spacetime dimension, but this membrane itself is probably not a truly fundamental object either. Strings and point particles are distinguished by the fact that a quantum theory which *does* regard them as fundamental objects can be consistently constructed along the lines that we have examined. No such theory of higher-dimensional objects appears to be possible; to investigate their properties, one must resort to less direct (and partly conjectural) methods, based on ideas such as the dualities that I have alluded to earlier.

At the beginning of this Tour, I undertook to discuss those central ideas which constitute our current understanding of the ways of nature. One may wonder, perhaps, whether the ideas of this chapter really do tell us anything about the ways of nature, or whether they are merely part of an elaborate mathematical game. To me, at least, the answer is far from obvious. As I emphasized at the beginning of the chapter, a view of the world that combines a quantum-mechanical theory of matter with classical general relativity as a theory of gravity provides an adequate means of accounting for all currently observed phenomena, but it is not a self-consistent view and is ultimately untenable. One great virtue of string theory and its generalizations is that it offers, in prospect at least, a fully self-consistent view of the world. It is the only known theory to do so—but this does not by any means show that it is the correct theory. A significant drawback (to my mind) is that this self-consistency is achieved at the expense of postulating an extravagant array of concepts and phenomena which not only have no basis in current observations, but may well be inaccessible to any conceivable experiments—but this does not by any means show that the theory is wrong.

It is not yet possible to say whether string (or M) theory is consistent with our present knowledge of particle physics. We have seen that string theory has, at the fundamental level, no adjustable parameters and that the allowed internal symmetries are tightly constrained by requirements of mathematical consistency. The same may well be true of M-theory. The absence of arbitrary choices to be made at this fundamental level is a second great virtue of the theory. However, there are many possible ways in which the extra dimensions may be compactified, and many different possibilities for the expectation values of *moduli* fields such as the dilaton Φ that appears in (15.247). Thus, the theory has many possible ‘vacua’, which is to say that there are many possibilities for the effective four-dimensional, low-energy theory which, we might hope, would reproduce the standard model. It is known that some of these vacua have some of the right features, such as the gauge group SU(3)×SU(2)×U(1), but the mechanism by which one particular vacuum might emerge as the one relevant to our universe (which must, in particular, involve the spontaneous breaking of supersymmetry) is not understood. A derivation from the first principles of string/M theory of some version of the standard model would, of course, provide a convincing vindication of the somewhat arcane ideas that we have touched on, but this is probably quite far off. On the other hand, if the brane-world picture is right, then some experimental indication of this might appear quite soon. Whether these ideas will prove to be the last word in the story of unified theories of the world, I do not know, but these remarks must be the last words of our Tour.

Exercises

15.1. Using a coordinate system in which the worldsheet metric has the form (15.29), show that the connection coefficients (2.50) are

$$\Gamma^a_{\ bc} = \tfrac{1}{2}\left[\delta^a_b\Omega_{,c} + \delta^a_c\Omega_{,b} - \eta_{bc}\Omega^{,a}\right]$$

where indices are raised and lowered using η^{ab} and η_{ab}, and that the Ricci tensor is

$$R_{ab} = -\tfrac{1}{2}\eta_{ab}\Omega^{,c}{}_{,c}.$$

Verify that $R_{ab} = \frac{1}{2}R\gamma_{ab}$ and note that since this is a tensor equation it is valid in any coordinate system. Show that $(-\gamma)^{1/2}R$, which appears in the two-dimensional version of the Einstein–Hilbert Lagrangian (4.16) is a total divergence, equal to $-\Omega^{,a}{}_{,a}$, which would not affect the equations of motion had we included it in the string action.

15.2. (a) Consider the configuration of an open string specified, relative to a particular frame of reference in spacetime, by $X^1(\tau,\sigma) = c\sigma$, where c is a constant, $\partial_\sigma X^0(\tau,\sigma) = 0$ and $X^\mu(\sigma,\tau) = 0$ for $\mu \geq 2$. Take the range of σ to be $0 \leq \sigma \leq \ell$. As viewed from this frame of reference, what is the length of the string, and what is its state of motion? Use the constraint (15.36) to find the value of $\partial_\tau X^0$ (assuming that this quantity is positive) and verify that *all* the components of T^{ab} vanish. Hence find the spacetime momentum (15.33) and verify that the mass per unit length is $1/2\pi\alpha'$. (Note that this configuration does not satisfy the boundary condition $\partial_\sigma X^\mu = 0$ at $\sigma = 0$ and $\sigma = \ell$. We must imagine its ends to be held in place by some external agency.)

(b) For an open string whose ends are not artificially held in place, use the boundary condition and the constraint to show that

$$\partial_\tau X_\mu(\tau,0)\partial_\tau X^\mu(\tau,0) = \partial_\tau X_\mu(\tau,\ell)\partial_\tau X^\mu(\tau,\ell) = 0$$

and deduce that the ends move with the speed of light.

15.3. Show that $\partial L_n/\partial\alpha^\mu_{m'} = -\alpha_{(n-m')\,\mu}$, where L_n is the Virasoro generator defined in (15.73). Now use the expression (15.76) for the Poisson bracket to show that

$$\{L_m, L_n\}_\mathrm{P} = \frac{\mathrm{i}}{2}\sum_{m'=-\infty}^{\infty} m'\left[\alpha_{(m-m')\,\mu}\alpha^\mu_{n+m'} - \alpha_{(n-m')\,\mu}\alpha^\mu_{m+m'}\right].$$

By making the change of summation variable $m' = m - m''$ in the first term and $m' = n - m''$ in the second, verify the result (15.77).

15.4. As given in (15.33), the spacetime momentum P^μ is an integral over a specific curve on the worldsheet, namely τ = constant in some particular coordinate system. We might wonder whether the value of P^μ depends on our choice of this curve. Show from the considerations of §15.2.5 that it does not.

15.5. At the end of §15.3.1, I asserted that there are infinitely many conserved quantities associated with the symmetry of conformal invariance. To understand what this means, consider the definition of a conserved current j^μ as one that satisfies the equation of continuity (3.40). Using complex coordinates in two dimensions, show that the equation of continuity takes the form

$$\bar{\partial} j_z(z,\bar{z}) + \partial j_{\bar{z}}(z,\bar{z}) = 0.$$

Now define the particular current

$$j_a^{(v)}(z,\bar{z}) = v^b(z,\bar{z}) T_{ab}(z,\bar{z})$$

where $v^a(z,\bar{z})$ is a vector field and $T_{ab}(z,\bar{z})$ is the energy–momentum tensor. Show that $j_a^{(v)}$ is conserved if $v^z(z)$ is any holomorphic function and $v^{\bar{z}}(\bar{z})$ is any antiholomorphic function. In terms of the coordinates σ and τ, there is a conserved 'charge' q corresponding to the conserved current $j_a^{(v)}$:

$$q^{(v)} = \int_0^{2\pi} \mathrm{d}\sigma\, j_\tau^{(v)} \qquad \partial_\tau q^{(v)} = \int_0^{2\pi} \mathrm{d}\sigma\, \partial_\tau j_\tau^{(v)} = \int_0^{2\pi} \mathrm{d}\sigma\, \partial_\tau \left(v^b T_{\tau b} \right) = 0.$$

By translating the first of these statements into the coordinates $z = \mathrm{e}^{\mathrm{i}(\tau-\sigma)}$ and $\bar{z} = \mathrm{e}^{\mathrm{i}(\tau+\sigma)}$, show that when $v^z = z^{n+1}$ and $v^{\bar{z}} = 0$, the charge $q^{(v)}$ is proportional to the Virasoro generator L_n. You might think, therefore, that L_n should obey the equation of motion $\partial_\tau L_n = -\mathrm{i}[L_n, H] = 0$, with the Hamiltonian $H = L_0 + \widetilde{L}_0$, but this is not consistent with the commutation relations (15.116) of the Virasoro algebra. The reason is that neither $q^{(v)}$ nor L_n is a *bona fide* Heisenberg-picture operator. The energy–momentum tensor T_{ab} *is* a Heisenberg-picture operator, whose time dependence is given by the equation of motion, but the components of the vector field v^a are just functions, which must be differentiated explicitly. Using the above expression for $\partial_\tau q^{(v)}$, verify that the commutation relations (15.116) do imply $\partial_\tau q^{(v)} = 0$ when used correctly. [This exercise may be quite tricky. You will need to carry out the coordinate transformations carefully, express the σ integral as a contour integral in the z plane and use Cauchy's theorem to extract answers in terms of the L_n.]

15.6. Use the connection coefficients of exercise 15.1 to show that the covariant divergence of a symmetric rank $\binom{0}{2}$ tensor such as the energy–momentum tensor is

$$\nabla^a T_{ab} = \partial^a T_{ab} - \tfrac{1}{2}\Omega_{,b} T_a^a$$

where indices are raised and lowered with the full metric.

15.7. The algebra of the ghost operators b_0 and c_0 can be represented by 2×2 matrices. Verify that the matrices

$$b_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad c_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

have the anticommutation relations $\{b_0, c_0\} = 1$ and $b_0^2 = c_0^2 = 0$ and that the basis vectors $|0_\mathrm{g}\rangle = \binom{1}{0}$ and $|1_\mathrm{g}\rangle = \binom{0}{1}$ have the properties exhibited in (15.164).

15.8. Use the (anti)commutation relations (15.99) and (15.160) to verify that the number operators defined in (15.199) obey the commutation relations

$$[\alpha^{\nu}_{-n}, N^{(X)}_m] = -\alpha^{\nu}_{-n}\,\delta_{m,n}$$

$$[b_{-n}, N^{(b)}_m] = -b_{-n}\delta_{m,n} \qquad [c_{-n}, N^{(c)}_m] = -c_{-n}\delta_{m,n}$$

when m and n are both positive. Hence show that $N^{(X)}_n$ counts the number of quanta created by the α^{ν}_{-n} and so on.

15.9. For an open string, the creation and annihilation operators of left- and right-moving modes are identical, so a general state at level 1 is $A_{-1}|0; \boldsymbol{k}\rangle$, where the creation operator is

$$A_{-1} = \epsilon_{\mu}\alpha^{\mu}_{-1} + \kappa b_{-1} + \lambda c_{-1}.$$

Use the (anti)commutation relations (15.205)–(15.207) to show that

$$Q\alpha^{\mu}_{-1} = \alpha^{\mu}_0 c_{-1} + \ldots \qquad Qb_{-1} = -\alpha^{\mu}_0\alpha_{-1\,\mu} + \ldots \qquad Qc_{-1} = \ldots$$

where '$\ldots$' means a collection of operators that produce zero when acting on $|0; \boldsymbol{k}\rangle$. Hence show that

(a) $A_{-1}|0; \boldsymbol{k}\rangle$ is a closed state if $k^{\mu}\epsilon_{\mu} = 0$ and $\kappa = 0$;

(b) the general form of an exact state at level 1 is

$$\left[-\kappa' k_{\mu}\alpha^{\mu}_{-1} + \epsilon'_{\mu}k^{\mu}c_{-1}\right]|0; \boldsymbol{k}\rangle$$

where κ' and ϵ'_{μ} are constants, which in general will be different from those used to construct a closed state.

Using these results, show that every closed state at level 1 is equivalent to a state of the form

$$\epsilon_{\mu}\alpha^{\mu}_{-1}|0; \boldsymbol{k}\rangle$$

where the polarization vector satisfies $k^{\mu}\epsilon_{\mu} = 0$, and that the two polarization vectors ϵ_{μ} and $\epsilon_{\mu} - \kappa k_{\mu}$ are equivalent for any constant κ.

15.10. In four spacetime dimensions, consider the frame of reference in which the momentum of a massless particle is $k^{\mu} = (k, 0, 0, k)$. According to (15.214), the polarization tensor $\epsilon^{\mu\nu}_{\mathrm{a}}$ is physically equivalent to a new polarization tensor

$$\epsilon'^{\mu\nu}_{\mathrm{a}} = \epsilon^{\mu\nu}_{\mathrm{a}} + k^{\mu}\xi^{\nu} - k^{\nu}\xi^{\mu}$$

where $\xi^{\mu} = \frac{1}{2}(\widetilde{\kappa}^{\mu} - \kappa^{\mu})$. Find the independent components of ξ^{μ} that are allowed by the constraint $k_{\mu}\xi^{\mu} = 0$ and show that they can be chosen so that $\epsilon'^{\mu\nu}_{\mathrm{a}}$ has just one independent, non-zero component $\epsilon'^{12}_{\mathrm{a}} = -\epsilon'^{21}_{\mathrm{a}}$.

Dropping the $'$ from this new tensor, show that its spatial components can be written as

$$\epsilon_{\mathrm{a}}^{ij} = a\hat{\epsilon}^{0ij\ell}k^{\ell}$$

where a is a constant and $\hat{\epsilon}^{\mu\nu\sigma\tau}$ is the Levi-Civita symbol. Investigate the transformation of $\epsilon_{\mathrm{a}}^{ij}$ under spatial rotations and spatial reflections ($\boldsymbol{x}' = -\boldsymbol{x}$, which also implies $\boldsymbol{k}' = -\boldsymbol{k}$). Verify that a transforms as a scalar under rotations, but changes sign under reflections. According to the classification of §7.3.5, a is a pseudoscalar. By analogy with an axial vector, which is a pseudovector, a particle whose polarization has this property is called an 'axion'.

15.11. In d spacetime dimensions, consider the frame of reference in which the momentum of a massless particle is $k^{\mu} = (k, 0, \ldots, 0, k)$. Show that the polarization vector of exercise 15.9 can be chosen so that it has $d - 2$ non-zero components in the spatial directions perpendicular to $\boldsymbol{k}$. In four dimensions, these are the two polarization states of a massless spin-1 vector boson.

15.12. On a Euclidean sphere of radius a, the line element can be written in terms of the usual polar angles as $\mathrm{d}s^2 = a^2\left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right)$. By exchanging θ for a coordinate ψ such that $\mathrm{d}\psi = \mathrm{d}\theta/\sin\theta$, show that the metric on the sphere can be written as $(\exp\Omega)\,\delta_{ab}$ with $\Omega = 2\ln(a\sin\theta)$. By adapting the results of exercise 15.1 to this Euclidean metric, show that the Euler characteristic of the sphere is $\chi = 2$. Note that this is independent of the radius a.

Some Snapshots of the Tour

Our tour having come to an end, readers may like to review some of its highlights with the aid of a few snapshots, which are provided on the following pages. The snapshots are intended to give an overall view of the logical structure of the principal theories we have visited and to summarize some of the important results.

Thank you for travelling with Unified Grand Tours; I hope that your journey has been a pleasant one and that these pages will continue to serve you well.

Snapshot of Geometry and Gravitation

Geometry

- The basic spacetime structure is a *differentiable manifold* on which smooth curves can be drawn and which can support differentiable functions. But *parallelism*, *angle* and *length* are not defined.

- *Tensors* may be defined either as intrinsic geometrical objects or as sets of components with definite transformation laws:

$$\boxed{T^{\mu'\ldots}{}_{\alpha'\ldots} = \left[\frac{\partial x^{\mu'}}{\partial x^{\mu}}\frac{\partial x^{\alpha}}{\partial x^{\alpha'}}\cdots\right] T^{\mu\ldots}{}_{\alpha\ldots}}$$

  Typical (contravariant) vector is the tangent vector $V^{\mu} = (\mathrm{d}/\mathrm{d}\lambda)x^{\mu}(\lambda)$ to a curve $x^{\mu}(\lambda)$.
  Typical one-form (covariant vector) is the gradient $\omega_{\mu} = \partial f/\partial x^{\mu}$ of a scalar function $f(x)$.

- The *affine connection* Γ defines parallel transport, yielding the *covariant derivative*

$$\boxed{\nabla_{\mu}T^{\alpha\ldots}{}_{\beta\ldots} = \partial_{\mu}T^{\alpha\ldots}{}_{\beta\ldots} + \Gamma^{\alpha}{}_{\lambda\mu}T^{\lambda\ldots}{}_{\beta\ldots} - \Gamma^{\lambda}{}_{\beta\mu}T^{\alpha\ldots}{}_{\lambda\ldots} + \ldots}$$

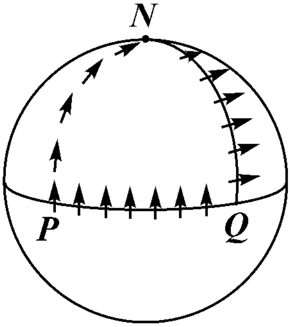
- *Curvature* of a manifold is defined (for a symmetric connection) by

$$[\nabla_{\mu}, \nabla_{\nu}]V^{\alpha} = R^{\alpha}{}_{\beta\mu\nu}V^{\beta}$$

  with the *Riemann tensor* given by

$$\boxed{R^{\alpha}{}_{\beta\mu\nu} = \Gamma^{\alpha}{}_{\beta\nu,\mu} - \Gamma^{\alpha}{}_{\beta\mu,\nu} + \Gamma^{\alpha}{}_{\sigma\mu}\Gamma^{\sigma}{}_{\beta\nu} - \Gamma^{\alpha}{}_{\sigma\nu}\Gamma^{\sigma}{}_{\beta\mu}}$$

  Roughly, it measures the extent to which the result of parallelly transporting a vector between two points depends on the route taken:



- The *Ricci tensor* is $\boxed{R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu}}$

- A *geodesic* is a self-parallel curve (generalization of straight line):

$$\boxed{\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu_{\ \nu\sigma} \frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} \frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} = 0}$$

for an affinely parametrized curve.

- The *metric tensor* $g_{\mu\nu}(x)$ defines

(a) distance along a curve $\boxed{\mathrm{d}s^2 = g_{\mu\nu}(x)\mathrm{d}x^\mu \mathrm{d}x^\nu}$

(b) scalar product of two vectors $\boxed{g_{\mu\nu} V^\mu V^\nu}$

(c) one-to-one correspondence between vectors and one-forms

$$\boxed{V_\mu = g_{\mu\nu} V^\nu} \qquad \boxed{V^\mu = g^{\mu\nu} V_\nu} \qquad \boxed{g^{\mu\sigma} g_{\sigma\nu} = \delta^\mu_\nu}$$

- The *metric connection* preserves lengths and angles under parallel transport. The requirement

$$\boxed{\nabla_\mu g_{\alpha\beta} = 0}$$

determines the metric connection coefficients (Christoffel symbols) as

$$\boxed{\Gamma^\mu_{\ \nu\sigma} = \tfrac{1}{2} g^{\mu\lambda} \left(g_{\nu\lambda,\sigma} + g_{\sigma\lambda,\nu} - g_{\nu\sigma,\lambda} \right)}$$

- The *Ricci curvature scalar* is $\boxed{R = g^{\mu\nu} R_{\mu\nu}}$

- A *vierbein* embeds local inertial coordinates y^a into a large-scale coordinate system x^μ:

$$\boxed{e^a_{\ \mu} = \frac{\partial y^a}{\partial x^\mu}} \qquad \boxed{e^a_{\ \mu} e^\mu_{\ b} = \delta^a_b} \qquad \boxed{e^a_{\ \mu} e^b_{\ \nu} \eta_{ab} = g_{\mu\nu}}$$

$$\boxed{e^\mu_{\ a} = \frac{\partial x^\mu}{\partial y^a}} \qquad \boxed{e^\mu_{\ a} e^a_{\ \nu} = \delta^\mu_\nu} \qquad \boxed{e^\mu_{\ a} e^\nu_{\ b} g_{\mu\nu} = \eta_{ab}}$$

Gravitation and cosmology

- *Equivalence principle*: at any point P, we can find a coordinate system such that

$$g_{\mu\nu}(P) = \eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

  (the Minkowski metric of special relativity) and $g_{\mu\nu,\sigma}(P) = 0$, but in general $g_{\mu\nu,\sigma\tau} \neq 0$.

- The presence of a *gravitational field* is indicated if it is impossible to find coordinates can such that $g_{\mu\nu} = \eta_{\mu\nu}$ everywhere.

- *Response of test particles to a gravitational field*: In the absence of non-gravitational forces, a particle follows a geodesic path. Connection terms in the geodesic equation are interpreted as gravitational forces. In the *Newtonian limit* of weak, static fields and slowly moving particles ($g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$)

$$\boxed{\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2} \approx -\frac{\partial V}{\partial x^i}} \qquad \boxed{V = \tfrac{1}{2}c^2 h_{00}}$$

  This implies the equality of inertial and gravitational masses.

- *Response of geometry to distribution of matter*: Einstein's *field equations*, which follow from a principle of least action, are

$$\boxed{R_{\mu\nu} - (\tfrac{1}{2}R + \Lambda)g_{\mu\nu} = \kappa T_{\mu\nu}} \qquad \boxed{\kappa = 8\pi G/c^4}$$

  In the Newtonian limit (with $\Lambda = 0$) we can deduce Poisson's equation

$$\boxed{\nabla^2 V \approx 4\pi G\rho/c^2}$$

  where ρ is the energy density of matter and ρ/c^2 is the equivalent mass density.

- For a perfect fluid, the stress–energy–momentum tensor is

$$\boxed{T^{\mu\nu} = c^{-2}(\rho + p)u^\mu u^\nu - pg^{\mu\nu}}$$

  where u^μ is the 4-velocity of a fluid element while ρ and p are the energy density and pressure measured in its rest frame.

- Quite generally, the stress tensor is 'covariantly conserved':

$$\boxed{\nabla_\nu T^{\mu\nu} = 0}$$

- *Schwarzschild's solution*, valid in the vacuum outside a spherically symmetric body of mass M, is

$$\boxed{c^2 \mathrm{d}\tau^2 = (1 - r_\mathrm{S}/r)c^2 \mathrm{d}t^2 - (1 - r_\mathrm{S}/r)^{-1} \mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)}$$

  where the *Schwarzschild radius* is $r_\mathrm{S} = 2GM/c^2$.

- The *Robertson–Walker* metric for a homogeneous, isotropic universe is

$$\boxed{\mathrm{d}\tau^2 = \mathrm{d}t^2 - a^2(t)\left[(1 - kr^2)^{-1}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)\right]}$$

  with $c = 1$. Cosmic time t is proper time for comoving observers. Spatial sections may be closed ($k = 1$), flat ($k = 0$) or open ($k = -1$). The stress tensor must have the perfect-fluid form.

- The field equations in Friedmann–Robertson–Walker models are

$$\boxed{3\left(\frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right) = \kappa\rho + \Lambda} \qquad \boxed{2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} = -\kappa p + \Lambda}$$

  These imply energy conservation in the form

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}(\rho a^3) = -p\frac{\mathrm{d}}{\mathrm{d}t}(a^3)}$$

- *Hubble's law* is

$$\boxed{\text{recessional velocity} = H(t) \times \text{distance}} \qquad \boxed{H(t) = \dot{a}(t)/a(t)}$$

- The *critical density* and *density ratio* are given by

$$\boxed{\rho_\mathrm{c}(t) = 3H^2(t)/\kappa} \qquad \boxed{\Omega(t) = \rho(t)/\rho_\mathrm{c}(t)}$$

  If $\Lambda = 0$, $\boxed{\Omega > 1 \Rightarrow k = 1; \quad \Omega = 1 \Rightarrow k = 0; \quad \Omega < 1 \Rightarrow k = -1}$

- In a flat, matter-dominated universe ($p = 0$): $a(t) \sim t^{2/3}$

- In a flat, radiation-dominated universe ($p = \frac{1}{3}\rho$): $a(t) \sim t^{1/2}$

Snapshot of Field Theory

Free fields

- *Klein–Gordon equation*: with the substitutions $E \to \mathrm{i}\partial_t$ and $\boldsymbol{p} \to -\mathrm{i}\nabla$,

$$E^2 = p^2 + m^2 \qquad \text{leads to} \qquad \boxed{(\Box + m^2)\phi(x) = 0}$$

- The *Dirac equation* for spin-$\frac{1}{2}$ particles is $\boxed{(\mathrm{i}\gamma^\mu \partial_\mu - m)\psi(x) = 0}$

  To reproduce the Klein–Gordon equation, the matrices γ^μ satisfy

$$\boxed{\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}}$$

- To accommodate negative energies, write general solutions as

$$\boxed{\phi(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})} \left[a(\boldsymbol{k})\mathrm{e}^{-\mathrm{i}k\cdot x} + c^\dagger(\boldsymbol{k})\mathrm{e}^{\mathrm{i}k\cdot x}\right]}$$

$$\boxed{\psi(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 2\omega(\boldsymbol{k})} \sum_s \left[b(\boldsymbol{k}, s)u(k, s)\mathrm{e}^{-\mathrm{i}k\cdot x} + d^\dagger(\boldsymbol{k}, s)v(k, s)\mathrm{e}^{\mathrm{i}k\cdot x}\right]}$$

  where $a(\boldsymbol{k})$ and $b(\boldsymbol{k}, s)$ annihilate particles, while $c^\dagger(\boldsymbol{k}, s)$ and $d^\dagger(\boldsymbol{k}, s)$ create antiparticles.

- For bosons, this follows from the canonical quantization procedure, using the commutator

$$\boxed{[\phi(\boldsymbol{x}, t), \Pi(\boldsymbol{x}', t)] = \mathrm{i}\delta(\boldsymbol{x} - \boldsymbol{x}')} \quad \boxed{\Pi(\boldsymbol{x}, t) = \delta S/\delta\dot{\phi}(\boldsymbol{x}, t) = \dot{\phi}^\dagger(\boldsymbol{x}, t)}$$

  from which we obtain

$$\boxed{[a(\boldsymbol{k}), a^\dagger(\boldsymbol{k}')] = (2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k} - \boldsymbol{k}') \;\Rightarrow \text{Bose–Einstein statistics}}$$

  For fermions, we must impose anticommutation relations

$$\boxed{\{b(\boldsymbol{k}, s), b^\dagger(\boldsymbol{k}', s')\} = (2\pi)^3 2\omega(\boldsymbol{k})\delta(\boldsymbol{k} - \boldsymbol{k}')\delta_{ss'} \;\Rightarrow \text{Fermi–Dirac statistics}}$$

Interacting fields

- *Asymptotic states*: Initial and final states of scattering/decay processes are created by free fields, $\phi_{\text{in}}(x)$ and $\phi_{\text{out}}(x)$. With adiabatic switching,

$$\begin{aligned}\phi(x) &\to \sqrt{Z}\phi_{\text{in}}(x) \qquad && t \to -\infty \\ &\to \sqrt{Z}\phi_{\text{out}}(x) \qquad && t \to +\infty.\end{aligned}$$

- Amplitudes $\langle k_1', \ldots; \text{out}|k_1, \ldots; \text{in}\rangle$ are related to vacuum expectation values of interacting fields by *reduction formulae*

$$\boxed{\langle \ldots; \text{out}| \ldots; \text{in}\rangle = \int \mathrm{d}x_1 \cdots \langle 0|T[\phi \cdots \phi^\dagger]|0\rangle \cdots}$$

- Time ordered products:

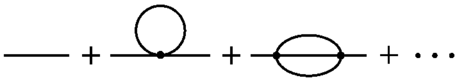
$$\boxed{T[\phi(x_1)\cdots\phi(x_N)] \quad \text{denotes latest-on-left ordering of fields.}}$$

When a product of fields is brought into time-ordered form, there is a factor (-1) for each interchange of two fermionic fields.

- Vacuum expectation values of time-ordered products have a path-integral representation

$$\boxed{\langle 0|T[\phi(x_1)\cdots\phi^\dagger(x_N)]|0\rangle = \int \mathcal{D}\phi\, \phi(x_1)\cdots\phi^*(x_N)\, \mathrm{e}^{\mathrm{i}S(\phi)}}$$

- *Perturbation theory*. Expansion in powers of coupling constants can be represented by *Feynman diagrams*, e.g. for $\lambda\phi^4$ theory:

$-\mathrm{i}\langle 0|T[\phi(x)\phi(y)]|0\rangle =$ 

Lines are unperturbed propagators; after Fourier transformation

$$\text{scalar:}\ \frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon} \qquad \text{spin-}\tfrac{1}{2}\text{:}\ \frac{\mathrm{i}(\not{p} + m)}{p^2 - m^2 + \mathrm{i}\epsilon}$$

4-momentum is conserved at each vertex and momentum flowing round internal loops is integrated.

- *Renormalization*: Re-expresses calculated results in terms of physically measurable masses, etc, which are modified by interactions. In *renormalizable* theories, infinities in Feynman integrals are removed by renormalization.
- *Running coupling constants*: Net effect of interactions involves combinations of coupling constants and energy–momentum-dependent Feynman integrals. Effective expansion parameter depends on energy. Also related to renormalization and characteristic length scales of physical processes.

Gauge fields

- Fundamental forces arise from communication between different points of spacetime.
- A wavefunction or field exists in an *internal space*; the collection of internal spaces at all spacetime points is a *fibre bundle*.
- Comparison of fields at different points requires a *gauge connection* to define parallel transport through the fibre bundle.
- The *gauge-covariant derivative* is

$$\boxed{D_\mu \psi = (\partial_\mu + \mathrm{i}gA_\mu)\psi}$$

- The *field strength tensor* is the gauge-field analogue of the Riemann curvature tensor:

$$\boxed{\begin{aligned} F_{\mu\nu} &= -(\mathrm{i}/g)[D_\mu, D_\nu] && \\ &= \partial_\mu A_\nu - \partial_\nu A_\mu && \text{(Abelian)} \\ &= \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}g[A_\mu, A_\nu] && \text{(non-Abelian).} \end{aligned}}$$

In the Abelian case, A_μ is the electromagnetic 4-vector potential and $F_{\mu\nu}$ is the Maxwell field-strength tensor, whose elements are the components of $\boldsymbol{E}$ and $\boldsymbol{B}$. In the non-Abelian case, $A_\mu = A_\mu^a T^a$, where T^a are the generator matrices of some representation of the gauge group.

- Gauge theories are invariant under *gauge transformations* (c.f. general coordinate transformations)

$$\boxed{\begin{aligned} \psi &\to U\psi \\ A_\mu &\to UA_\mu U^{-1} + (\mathrm{i}/g)(\partial_\mu U)U^{-1} \end{aligned}}$$

  In the Abelian case, $U = \mathrm{e}^{\mathrm{i}\theta(x)}$ is a phase transformation; in the non-Abelian case, ψ is a multiplet of fields, and U is a matrix which rearranges its components.

- The gauge-covariant derivative $D_\mu\psi$ transforms in the same way as ψ, namely $D_\mu\psi \to UD_\mu\psi$.

- The gauge-invariant action for a theory of gauge fields and fermions has the form

$$\boxed{S = \int \mathrm{d}^4x \left[-\tfrac{1}{4}\,\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu}) + \bar{\psi}(\mathrm{i}\gamma^\mu D_\mu - m)\psi\right]}$$

- *Gauge-boson masses*: The quantity $M^2 A_\mu A^\mu$ is not gauge invariant. Masses can be introduced in a gauge-invariant manner through a *Higgs scalar field*, with the action

$$\boxed{S_{\mathrm{Higgs}} = \int \mathrm{d}^4x \left[(D_\mu\phi)^\dagger(D^\mu\phi) - V(\phi^\dagger\phi)\right]}$$

  which acquires a non-zero vacuum expectation value.

- When the left- and right-handed chiral components of ψ have different gauge transformation laws, the quantity $m\bar{\psi}\psi = m(\bar{\psi}_{\mathrm{L}}\psi_{\mathrm{R}} + \bar{\psi}_{\mathrm{R}}\psi_{\mathrm{L}})$ is not gauge invariant either. In the standard electroweak theory, fermion masses can be generated in a gauge-invariant manner through Yukawa couplings to Higgs fields. For electrons, for example:

$$\boxed{\Delta\mathcal{L} = -f(\bar{\ell}_{\mathrm{e}}\phi e_{\mathrm{R}} + \bar{e}_{\mathrm{R}}\phi^\dagger\ell_{\mathrm{e}})}$$

Snapshot of Statistical Mechanics

- *Classical ergodic theory*: Ensemble average of a dynamical quantity is

$$\boxed{\bar{A}(t) = \int \mathrm{d}^{6N}X\, \rho(X,t)A(X)}$$

where X is a point in phase space, and $\rho(X)$ is the probability density.

- The Liouville equation for the probability density is

$$\boxed{\frac{\partial \rho}{\partial t} = \{H, \rho\}_{\mathrm{P}}}$$

- To describe thermal equilibrium, we need $\partial\rho/\partial t = 0$, so ρ depends on X only through conserved quantities.

- *Closed, isoenergetic system*: An isolated system, with fixed energy E, particle number N and volume V is described by the *microcanonical ensemble*

$$\boxed{\rho_{\mathrm{micro}}(X,E) = \delta[H_N(X) - E]/\Sigma(E,N,V)}$$

$$\boxed{\Sigma(E,N,V) = \int \mathrm{d}^{6N}X\, \delta[H_N(X) - E]}$$

Ergodic system ↔ microcanonical average = long-time average
↔ ρ_{micro} is the unique time-independent distribution.

Relation to thermodynamics:

$$\boxed{\text{Entropy } S(E,N,V) = k_{\mathrm{B}} \ln\left[\Sigma(E,N,V)/h^{3N}N!\right]}$$

- *Closed isothermal system*: A system with fixed N and V in equilibrium with a heat bath at temperature T is described by the *canonical ensemble*. Statistical independence of non-interacting systems implies the probability density

$$\boxed{\rho_{\mathrm{can}}(X,\beta) = \mathrm{e}^{-\beta H_N(X)}\left[\int \mathrm{d}^{6N}X\, \mathrm{e}^{-\beta H_N(X)}\right]^{-1} \qquad \beta = 1/k_{\mathrm{B}}T}$$

The canonical partition function is

$$\boxed{Z_{\text{can}}(\beta, N, V) = (h^{3N} N!)^{-1} \int \mathrm{d}^{6N} X \, \mathrm{e}^{-\beta H_N(X)}}$$

Relation to thermodynamics:

$$\boxed{\text{Helmholtz free energy } F(\beta, N, V) = -k_{\text{B}} T \ln Z_{\text{can}}(\beta, N, V)}$$

- The canonical and microcanonical ensembles are related by a Laplace transform

$$\boxed{Z_{\text{can}}(\beta, N, V) = (h^{3N} N!)^{-1} \int \mathrm{d}E \, \mathrm{e}^{-\beta E} \Sigma(E, N, V)}$$

and in the thermodynamic limit (regarding internal energy U as equivalent to E), the thermodynamic potentials are related by a Legendre transform

$$\boxed{F = U - TS}$$

- *Open isothermal system:* A system of fixed volume in equilibrium with a heat bath and particle reservoir at temperature T and chemical potential μ is described by the *grand canonical ensemble*. Poisson distribution of particle numbers implies

$$\boxed{\rho_{\text{gr}}(X, N) = \mathrm{e}^{\beta\mu N} \mathrm{e}^{-\beta H_N(X)} \left[h^{3N} N! Z_{\text{gr}}\right]^{-1}}$$

$$\boxed{Z_{\text{gr}}(\beta, \mu, V) = \sum_N \mathrm{e}^{\beta\mu N} Z_{\text{can}}(\beta, N, V)}$$

The relation between canonical and grand canonical partition functions is a Laplace transform. Relation to thermodynamics:

$$\boxed{\text{Grand potential } \Omega(\beta, \mu, V) = -k_{\text{B}} T \ln Z_{\text{gr}}(\beta, \mu, V)}$$

In the thermodynamic limit, canonical and grand canonical potentials are related by a Legendre transform $\boxed{\Omega = F - \mu N}$

- Extensivity of the entropy implies

$$\boxed{TS = U + pV - \mu N \qquad \Omega = -pV}$$

Quantum statistical mechanics

- Expectation value of a dynamical quantity is

$$\boxed{\bar{A}(t) = \mathrm{Tr}[\,\hat{\rho}(t)\hat{A}\,]}$$

- Given a basis of Schrödinger-picture states $|\psi_n(t)\rangle$, the *density operator* is

$$\boxed{\hat{\rho}(t) = \sum_n |\psi_n(t)\rangle P_n \langle\psi_n(t)|}$$

  where P_n is the probability of finding the system in the nth quantum state. The expectation value is an average over both the quantum indeterminacy and our ignorance of the actual quantum state.

- Time evolution of the density operator is governed by the quantum Liouville equation

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}(t) = \frac{\mathrm{i}}{\hbar}\left[\hat{\rho}(t), \hat{H}\right]}$$

  and a system in thermal equilibrium is again described by a stationary density operator.

- Canonical ensemble:

$$\boxed{\hat{\rho}_{\mathrm{can}} = \mathrm{e}^{-\beta\hat{H}_N} Z_{\mathrm{can}}^{-1}} \qquad \boxed{Z_{\mathrm{can}}(\beta, N, V) = \mathrm{Tr}\ \mathrm{e}^{-\beta\hat{H}_N}}$$

- Grand canonical ensemble

$$\boxed{Z_{\mathrm{gr}}(\beta, \mu, V) = \sum_N \mathrm{e}^{\beta\mu N} Z_{\mathrm{can}}(\beta, N, V)}$$

  or, using second quantization

$$\boxed{\hat{\rho}_{\mathrm{gr}} = \mathrm{e}^{\beta\mu\hat{N} - \beta\hat{H}}} \qquad \boxed{Z_{\mathrm{gr}}(\beta, \mu, V) = \mathrm{Tr}\ \hat{\rho}_{\mathrm{gr}}}$$

- The grand canonical ensemble for a system of particles is equivalent to the canonical ensemble for an underlying system of quantum fields.

Field theories at finite temperature

- Field operators depending on an *imaginary time* τ $(0 \leq \tau \leq \beta)$ are defined by

$$\boxed{\hat{\phi}(\boldsymbol{x},\tau) = \mathrm{e}^{\tau\hat{H}}\hat{\phi}(\boldsymbol{x})\mathrm{e}^{-\tau\hat{H}} \qquad \hat{\phi}^\dagger(\boldsymbol{x},\tau) = \mathrm{e}^{\tau\hat{H}}\hat{\phi}^\dagger(\boldsymbol{x})\mathrm{e}^{-\tau\hat{H}}}$$

- Expectation values have the path-integral representation

$$\boxed{\mathrm{Tr}\left\{\hat{\rho}\, T_\tau\Big[\hat{\phi}(x_1)\cdots\hat{\phi}^\dagger(x_N)\Big]\right\} = Z_{\mathrm{gr}}^{-1}\int \mathcal{D}\phi\,\phi(x_1)\cdots\phi^*(x_N)\mathrm{e}^{-S_\beta(\phi)}}$$

and, for example, in $\lambda\phi^4$ theory

$$\boxed{S_\beta(\phi) = \int_0^\beta \mathrm{d}\tau\int \mathrm{d}^dx\left[\frac{\partial\phi^*}{\partial\tau}\frac{\partial\phi}{\partial\tau} + \nabla\phi^*\cdot\nabla\phi + m^2\phi^*\phi + \frac{\lambda}{4}(\phi^*\phi)^2\right]}$$

equivalent to a classical theory in $(d+1)$ Euclidean dimensions, of finite extent β in the extra dimension and with periodic boundary conditions (antiperiodic for fermions).

- For bosons, the imaginary-time propagator is

$$\boxed{G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau') = \mathrm{Tr}\Big[\hat{\rho}\, T_\tau[\hat{\phi}(\boldsymbol{x},\tau)\hat{\phi}^\dagger(\boldsymbol{x}',\tau')]\Big]}$$

It is periodic in imaginary time:

$$\boxed{G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau'\pm\beta) = G(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau')}$$

For a non-interacting theory, or at the lowest order of perturbation theory, its Fourier transform is

$$\boxed{G_0(\boldsymbol{x}-\boldsymbol{x}',\tau-\tau') = \beta^{-1}\int\frac{\mathrm{d}^dk}{(2\pi)^d}\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{x}-\boldsymbol{x}')}\sum_n \mathrm{e}^{\mathrm{i}\omega_n(\tau-\tau')}\widetilde{G}_0(\boldsymbol{k},n)}$$

$$\boxed{\widetilde{G}_0(\boldsymbol{k},n) = [k^2+\omega_n^2+m^2]^{-1}} \qquad \boxed{\omega_n = 2\pi n/\beta}$$

The ω_n are called the *Matsubara frequencies*.

- For fermions, the propagator is antiperiodic, and the Matsubara frequencies are $\omega_n = 2\pi(n+\frac{1}{2})/\beta$.

Snapshot of Bosonic String Theory

The classical string

- A string moving through Minkowski spacetime traces out a worldsheet, whose points we label with coordinates (τ, σ). Its action is

$$\boxed{S = -\frac{1}{4\pi\alpha'}\int_{-\infty}^{\infty} \mathrm{d}\tau \int_{0}^{\ell} \mathrm{d}\sigma\ (-\gamma)^{1/2}\gamma^{ab}\partial_a X_\mu \partial_b X^\mu}$$

where $X^\mu(\tau, \sigma)$ are the spacetime coordinates of the point (τ, σ). The worldsheet metric $\gamma_{ab}(\tau, \sigma)$ is independent of the spacetime metric.

- The Euler–Lagrange equation obtained by varying X^μ is

$$\boxed{\gamma^{ab}\nabla_a\nabla_b X^\mu = 0 \qquad \text{the equation of motion for the string}}$$

By varying γ_{ab} we get the constraint equation $\boxed{T^{ab}(\tau, \sigma) = 0}$ with

$$\boxed{T^{ab} = -\frac{1}{\alpha'}\left[\partial^a X_\mu \partial^b X^\mu - \tfrac{1}{2}\gamma^{ab}\partial_c X_\mu \partial^c X^\mu\right] \quad \begin{array}{l}\text{energy–momentum}\\ \text{tensor of the world-}\\ \text{sheet field theory}\end{array}}$$

- The action has a gauge symmetry, consisting of

$$\boxed{\begin{array}{ll} \begin{array}{l}\tau \to \tau'(\tau, \sigma)\\ \sigma \to \sigma'(\tau, \sigma)\end{array} & \text{reparametrization (or diffeomorphism) invariance}\\ \gamma_{ab}(\tau, \sigma) \to \exp[\omega(\tau, \sigma)]\gamma_{ab}(\tau, \sigma) & \text{Weyl invariance}\end{array}}$$

Using a combination of these transformations, we can 'fix the gauge', bringing the worldsheet metric to the form η_{ab}. After gauge fixing, the equation of motion and the energy–momentum tensor are

$$\boxed{\left[\partial_\tau^2 - \partial_\sigma^2\right] X^\mu = 0} \qquad \boxed{T^{ab} = -\frac{1}{\alpha'}\left[\partial^a X_\mu \partial^b X^\mu - \tfrac{1}{2}\eta^{ab}\partial_c X_\mu \partial^c X^\mu\right]}$$

- Useful coordinates on the worldsheet are defined by

$$\boxed{w = \sigma - \tau \qquad \bar{w} = \sigma + \tau} \quad \text{and} \quad \boxed{z = \mathrm{e}^{-\mathrm{i}w} \qquad \bar{z} = \mathrm{e}^{\mathrm{i}\bar{w}}}$$

On a *Euclidean worldsheet*, where τ is imaginary, the gauge-fixed line element is $\mathrm{d}s^2 = \mathrm{d}w\mathrm{d}\bar{w}$ and $\bar{w}$ and $\bar{z}$ are the complex conjugates of w and z. With $\partial = \partial/\partial w$ and $\bar{\partial} = \partial/\partial \bar{w}$, the Euclidean action is

$$\boxed{S_{\mathrm{E}} = -\frac{1}{2\pi\alpha'}\int \mathrm{d}w\mathrm{d}\bar{w}\,\partial X_\mu \bar{\partial} X^\mu}$$

It is invariant under *conformal transformations*, which are special combinations of a diffeomorphism and a Weyl transformation

$$\boxed{w' = f(w) \qquad \bar{w}' = \bar{f}(\bar{w}) \qquad \mathrm{d}s'^2 = \left|\frac{\mathrm{d}f}{\mathrm{d}w}\right|^2 \mathrm{d}s^2}$$

The transformation $z = f(w) = \mathrm{e}^{-\mathrm{i}w}$ is a particular case.

- The general solution to the equation of motion for a closed string is

$$\boxed{X^\mu(\tau,\sigma) = x^\mu + \alpha' p^\mu \tau + \mathrm{i}\left(\frac{\alpha'}{2}\right)^{1/2} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \frac{1}{n}\left[\alpha_n^\mu \mathrm{e}^{\mathrm{i}nw} + \widetilde{\alpha}_n^\mu \mathrm{e}^{-\mathrm{i}n\bar{w}}\right]}$$

where x^μ are the coordinates of the centre of mass and p^μ the spacetime momentum. The energy–momentum tensor has components

$$\boxed{T \equiv T_{ww} = \sum_{n=-\infty}^{\infty} L_n \mathrm{e}^{\mathrm{i}w} \qquad \widetilde{T} \equiv T_{\bar{w}\bar{w}} = \sum_{n=-\infty}^{\infty} \widetilde{L}_n \mathrm{e}^{-\mathrm{i}\bar{w}}}$$

The coefficients L_n and $\widetilde{L}_n$, which are the generators of conformal transformations, constitute the *Virasoro algebra*. They are given by

$$\boxed{L_n = -\tfrac{1}{2}\sum_{m=-\infty}^{\infty} \alpha_{m\,\mu}\alpha_{n-m}^\mu \qquad \widetilde{L}_n = -\tfrac{1}{2}\sum_{m=-\infty}^{\infty} \widetilde{\alpha}_{m\,\mu}\widetilde{\alpha}_{n-m}^\mu}$$

with $\alpha_0^\mu = \widetilde{\alpha}_0^\mu = (\alpha'/2)^{1/2} p^\mu$. The expansion coefficients α_n^μ and L_n for right-moving modes have the Poisson-bracket relations

$$\boxed{\{\alpha_m^\mu, \alpha_n^\nu\}_{\mathrm{P}} = \mathrm{i}m\eta^{\mu\nu}\delta_{m,-n} \qquad \{L_m, L_n\}_{\mathrm{P}} = -\mathrm{i}(m-n)L_{m+n}}$$

and for left-moving modes the $\widetilde{\alpha}_n^\mu$ and $\widetilde{L}_n$ obey the same relations. The constraint $T_{ab} = 0$ implies in particular that $L_0 = \widetilde{L}_0 = 0$ and this gives the mass of the string as

$$\boxed{M^2 \equiv p_\mu p^\mu = -\frac{4}{\alpha'}\sum_{n=1}^{\infty} \alpha_{-n\,\mu}\alpha_n^\mu = -\frac{4}{\alpha'}\sum_{n=1}^{\infty} \widetilde{\alpha}_{-n\,\mu}\widetilde{\alpha}_n^\mu}$$

The quantum string

- To quantize the classical string, we promote Poisson brackets to commutators:

$$\boxed{[\alpha^\mu_m, \alpha^\nu_n] = [\widetilde{\alpha}^\mu_m, \widetilde{\alpha}^\nu_n] = -m\eta^{\mu\nu}\delta_{m,-n} \qquad [x^\mu, p^\nu] = -\mathrm{i}\eta^{\mu\nu}}$$

  The modes of vibration of the string constitute an infinite set of harmonic oscillators. For the nth oscillator, quanta of energy are created by α^μ_{-n} (with $n > 0$) and annihilated by α^μ_n. From these commutation relations, we deduce those of the quantum Virasoro algebra

$$\boxed{[L_m, L_n] = \frac{m(m^2-1)}{12}\, c\, \delta_{m,-n} + (m-n)L_{m+n}}$$

  For a general conformal field theory, c is the *central charge*; for the X^μ field theory, c is equal to the number of spacetime dimensions d. Compared with the classical Poisson-bracket algebra, the extra term containing c arises from a *conformal anomaly*.

- Gauge fixing in the quantum theory requires the introduction of *Fadeev–Popov ghosts* b and c, resulting in an effective action

$$\boxed{S_{\mathrm{E}} = -\frac{1}{2\pi\alpha'}\int \mathrm{d}z\mathrm{d}\bar{z}\left[\partial X_\mu \bar{\partial} X^\mu - \alpha'\left(b\,\bar{\partial}c + \widetilde{b}\,\partial\widetilde{c}\right)\right]}$$

  but this is valid only when the central charge $c = d - 26$ of the combined theory vanishes, leading to a *critical spacetime dimension* $d = 26$ for the bosonic string.

- The gauge-fixed theory has a residual *BRST symmetry*, generated by the BRST charge Q, which is *nilpotent*: $Q^2 = 0$. The full Hilbert space for the X^μ + ghost theory can be constructed by acting with arbitrary combinations of creation operators α^μ_{-n}, b_{-n} and c_{-n} on the ground state $|\Omega\rangle$, but the states thus formed contain many unphysical gauge degrees of freedom. Physical states are those obeying the conditions

$$\boxed{Q|\psi\rangle = b_0|\psi\rangle = \widetilde{b}_0|\psi\rangle = 0 \qquad \text{conditions for physical states}}$$

  with the proviso that any two such states which differ by an *exact* vector, of the form $Q|\chi\rangle$, are physically equivalent. These conditions also imply that $L_0|\psi\rangle = \widetilde{L}_0|\psi\rangle = 0$, which determines the mass of each physically allowed state as $\boxed{M^2 = \frac{4}{\alpha'}(N-1)}$ where the level N is

$$\boxed{N = \sum_{n=1}^{\infty} n\left(N_n^{(X)} + N_n^{(b)} + N_n^{(c)}\right) = \sum_{n=1}^{\infty} n\left(\widetilde{N}_n^{(X)} + \widetilde{N}_n^{(b)} + \widetilde{N}_n^{(c)}\right)}$$

Here $N_n^{(X)}$ $(\widetilde{N}_n^{(X)})$ is the number of quanta in the nth right-moving (left-moving) mode of vibration and similarly for the ghost oscillators. The contributions of left- and right-moving modes are constrained to be equal.

- The lightest state of a closed string, with $N = 0$ is a *tachyon*, with $M^2 = -4/\alpha'$. Each massless state, with $N = 1$, is gauge-equivalent to a state of the form

$$\boxed{\epsilon_{\mu\nu}(k)\alpha_{-1}^{\mu}\widetilde{\alpha}_{-1}^{\nu}|0; \boldsymbol{k}\rangle \qquad \text{general massless state}}$$

The states for which $\epsilon_{\mu\nu}$ is symmetric and traceless ($\epsilon_{\mu}^{\mu} = 0$) can be identified as *graviton* states.

- A worldsheet with an 'end' corresponding to an incoming or outgoing particle with momentum $\boldsymbol{k}$ at the point with coordinates $(z, \bar{z})$ can be created by acting on $|\Omega\rangle$ with a *vertex operator* $\mathcal{V}(z, \bar{z}; \boldsymbol{k})$. In particular, the vertex operator for a graviton is

$$\boxed{\mathcal{V}_{\mathrm{g}}(z, \bar{z}; \boldsymbol{k}) = -\frac{2}{\alpha'}\epsilon_{\mu\nu} : \partial X^{\mu}(z)\bar{\partial}X^{\nu}(\bar{z}) \exp\left[-\mathrm{i}k \cdot X(z, \bar{z})\right] :}$$

- The Euclidean worldsheet action for a string moving through a curved spacetime is

$$\boxed{S_{\mathrm{E}} = -\frac{1}{2\pi\alpha'}\int \mathrm{d}z\mathrm{d}\bar{z}\, g_{\mu\nu}(X)\partial X^{\mu}\bar{\partial}X^{\nu}}$$

where $g_{\mu\nu}(X)$ is the spacetime metric. A small change in the spacetime metric, say $h_{\mu\nu}(X) = \epsilon_{\mu\nu}(k)\mathrm{e}^{-\mathrm{i}k\cdot X}$, leads to a change in the action

$$\boxed{\delta S_{\mathrm{E}}(\boldsymbol{k}) = \frac{1}{4\pi}\int \mathrm{d}z\mathrm{d}\bar{z}\, \mathcal{V}_{\mathrm{g}}(z, \bar{z}; \boldsymbol{k})}$$

This indicates that changes in spacetime geometry are equivalent to the emission and absorption of gravitons.

Appendix A

Some Mathematical Notes

This appendix contains a miscellaneous assortment of mathematical ideas and results. Some of them will be needed by readers who wish to verify the details of calculations presented in the main text; others are intended to indicate briefly how concepts that I have used in an informal way can be formulated more precisely. The topics are arranged more or less in the order in which they arise in the main text.

A.1 Delta Functions and Functional Differentiation

The *Kronecker delta* symbol, written as δ_{ij}, δ^{ij} or δ^i_j according to context, is defined to equal 1 if $i = j$ and 0 otherwise. It is mainly useful when we are dealing with summations, say of a set of quantities $\{f_i\}$, and it obviously has the property

$$\sum_i \delta_{ij} f_i = f_j . \tag{A.1}$$

The *Dirac delta function* is a generalization of the Kronecker δ which allows us to deal with integrals in the same way. The function (known in rigorous mathematics as a *distribution*) $\delta(x - x_0)$ is equal to zero unless $x = x_0$, when it is infinite. The infinite value becomes meaningful when the delta function appears inside an integral, and the defining property of $\delta(x - x_0)$ is that, for any sufficiently smooth function $f(x)$,

$$\int_a^b \delta(x - x_0)\, f(x)\, \mathrm{d}x = \begin{cases} f(x_0) & \text{if } a < x_0 < b \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

This can be understood in terms of the Riemann definition of the integral, according to which we divide the interval $[a, b]$ into N segments of length $\Delta x = (b - a)/N$ and take the limit $N \to \infty$. If x_0 lies in the jth segment, then the integral (A.2) can be represented as

$$\lim_{N\to\infty}\sum_i \Delta x\,\frac{\delta_{ij}}{\Delta x}\,f(x_i)=f(x_0) \tag{A.3}$$

and so $\delta(x-x_0)$ is the limit as $\Delta x\to 0$ of $\delta_{ij}/\Delta x$.

Consider a function $F(\{f_i\})$ which depends on all the f_i. If we make a small change Δf_i in each of these variables, then the first-order change in F is given by

$$\Delta F=\sum_i \Delta f_i\frac{\partial F}{\partial f_i}. \tag{A.4}$$

A *functional* $F[f]$, whose argument is a continuous function f, is a quantity whose value depends on infinitely many variables $f(x)$; there is one of these variables for each value of x. The electromagnetic action (3.53), for example, is a functional $S[A]$ whose arguments are the functions $A_\mu(x)$. We may ask how F changes when we make a small change $\Delta f(x)$ in the function $f(x)$. An equation analogous to (A.4), namely

$$\Delta F=\int \mathrm{d}x\,\Delta f(x)\,\frac{\delta F}{\delta f(x)} \tag{A.5}$$

defines the *functional derivative* $\delta F/\delta f(x)$, which is a generalization of the partial derivative $\partial F/\partial f_i$. The derivation of the Euler–Lagrange equations, which we first met in §3.1, is an example of functional differentiation, and Newton's law (3.1) might be written as $\delta S/\delta x(t)=0$. Quite often, F will be defined (like the action) as an integral whose integrand contains $f(x)$. In that case, F is not itself a function of x, but the functional derivative $\delta F/\delta f(x)$ *is* a function of x. On the other hand, we might take, for example, $F=f(y)$, which means that F really depends only on the single variable $f(y)$, which is the value of f at the particular point y. The functional derivative with respect to the variable $f(x)$ will be zero unless $x=y$. In fact, the definition (A.5) shows that $\delta f(y)/\delta f(x)=\delta(x-y)$, because

$$\Delta F=\int \mathrm{d}x\,\Delta f(x)\,\delta(x-y)=\Delta f(y). \tag{A.6}$$

The delta function in (A.2) can be thought of as imposing the constraint that $x=x_0$. Sometimes, we may wish instead to impose the constraint $g(x)=g_0$, where g is some function. In (7.9), for example, we use $g(k^0)=(k^0)^2$ and $g_0=\omega^2(\boldsymbol{k})$. This can be done by changing the integral over x to an integral over g:

$$\int \delta\left(g(x)-g_0\right)f(x)\mathrm{d}x=\int \mathrm{d}g\left(\frac{\mathrm{d}g}{\mathrm{d}x}\right)^{-1}\delta(g-g_0)f\left(x(g)\right)=\frac{f(x_0)}{g'(x_0)} \tag{A.7}$$

where x_0 is the point at which $g(x_0) = g_0$. If there are several such points, then the integral is the sum of values of f/g' at these points. We can therefore write

$$\delta\Big(g(x) - g_0\Big) = \sum_i \left(\frac{\mathrm{d}g}{\mathrm{d}x}\right)^{-1} \delta(x - x_i) \tag{A.8}$$

where x_i are the points at which $g(x_i) = g_0$.

In this book, I use $\delta(\boldsymbol{x} - \boldsymbol{x}')$ to stand for the product of delta functions $\delta(x - x')\delta(y - y')\delta(z - z')$.

A useful representation of the Dirac delta function is provided by the theory of Fourier transforms. If $f(x)$ is sufficiently well behaved, it can be expressed as

$$f(x) = \int_{-\infty}^{\infty} \mathrm{d}k \, \widetilde{f}(k) \, \mathrm{e}^{\mathrm{i}kx} \tag{A.9}$$

where

$$\widetilde{f}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}x \, f(x) \, \mathrm{e}^{-\mathrm{i}kx}. \tag{A.10}$$

By substituting (A.10) into (A.9)—or the other way round—and comparing the result with (A.2), we see that the delta function can be represented by

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \, \mathrm{e}^{\pm \mathrm{i}k(x - x')}. \tag{A.11}$$

Under suitable conditions, other orthogonal functions may be used in place of the exponential.

The *Heaviside step function* $\theta(x - x_0)$ is defined to equal 0 for $x < x_0$ and 1 for $x > x_0$. It is usually not necessary to specify its value at $x = x_0$. A little thought will show that $\mathrm{d}\theta(x - x_0)/\mathrm{d}x = \delta(x - x_0)$.

A.2 The Levi-Civita Tensor Density

The symbol $\epsilon^{\mu\nu\sigma\tau}$, in which each index can take the values 0, 1, 2 or 3, is defined to be $+1$ when $(\mu, \nu, \sigma, \tau) = (0, 1, 2, 3)$ and to be antisymmetric under the interchange of *any* pair of indices: $\epsilon^{\mu\nu\sigma\tau} = -\epsilon^{\nu\mu\sigma\tau} = \epsilon^{\nu\sigma\mu\tau}$, etc. It follows from this definition that $\epsilon^{\mu\nu\sigma\tau}$ is $+1$ when (μ, ν, σ, τ) is an even permutation of $(0, 1, 2, 3)$, -1 for an odd permutation and zero otherwise. Any totally antisymmetric tensor has only one independent component and is therefore proportional to ϵ. An ϵ symbol can be defined in any number of dimensions, d, by giving it d indices. The ϵ symbol can be made into a tensor-like quantity, called the *Levi-Civita tensor density*, by specifying its transformation properties. Suppose that its components have the values specified above in a particular coordinate system. In a new system, let

$$\hat{\epsilon}^{\mu'\nu'\sigma'\tau'} = \Lambda^{\mu'}{}_{\mu} \Lambda^{\nu'}{}_{\nu} \Lambda^{\sigma'}{}_{\sigma} \Lambda^{\tau'}{}_{\tau} \epsilon^{\mu\nu\sigma\tau}. \tag{A.12}$$

Clearly, $\hat{\epsilon}^{\mu'\nu'\sigma'\tau'}$ is also totally antisymmetric and therefore proportional to $\epsilon^{\mu'\nu'\sigma'\tau'}$. Furthermore, we have

$$\hat{\epsilon}^{0123} = \Lambda^0{}_\mu \Lambda^1{}_\nu \Lambda^2{}_\sigma \Lambda^3{}_\tau \epsilon^{\mu\nu\sigma\tau} = \det(\Lambda^{\mu'}{}_\mu) \tag{A.13}$$

since the sum of products with alternating signs is just the rule for forming the determinant. Thus, the ϵ symbol itself, which has exactly the same set of values in every coordinate system, obeys the transformation law

$$\epsilon^{\mu'\nu'\sigma'\tau'} = [\det(\Lambda^{\mu'}{}_\mu)]^{-1} \Lambda^{\mu'}{}_\mu \Lambda^{\nu'}{}_\nu \Lambda^{\sigma'}{}_\sigma \Lambda^{\tau'}{}_\tau \epsilon^{\mu\nu\sigma\tau}. \tag{A.14}$$

An object which transforms like a tensor, but with an extra factor of $[\det(\Lambda)]^n$ is called a *tensor density of weight n*, so ϵ is a tensor density of weight -1.

The metric determinant g can be written as

$$g = \det(g_{\mu\nu}) = \frac{1}{4!}\epsilon^{\mu\nu\sigma\tau}\epsilon^{\alpha\beta\gamma\delta} g_{\mu\alpha} g_{\nu\beta} g_{\sigma\gamma} g_{\tau\delta}. \tag{A.15}$$

This expression is 1/4! times a sum of 4! terms, each of them equal to $\epsilon^{\alpha\beta\gamma\delta} g_{0\alpha} g_{1\beta} g_{2\gamma} g_{3\delta}$, which is equal to $\det(g_{\mu\nu})$. Since each of the ϵ symbols in (A.15) transforms with a factor of $[\det(\Lambda)]^{-1}$, this determinant is a scalar density of weight -2.

It is convenient to define the covariant Levi-Civita symbol $\epsilon_{\mu\nu\sigma\tau}$ to have exactly the same values as $\epsilon^{\mu\nu\sigma\tau}$. In a manifold with a metric, this is not necessarily the quantity that we obtain by lowering the indices of $\epsilon^{\mu\nu\sigma\tau}$. In fact the same argument that gave us the transformation law (A.14) shows that

$$g_{\mu\alpha} g_{\nu\beta} g_{\sigma\gamma} g_{\tau\delta} \epsilon^{\alpha\beta\gamma\delta} = g\epsilon_{\mu\nu\sigma\tau}. \tag{A.16}$$

The left-hand side of this equation is a tensor density of weight -1, so $\epsilon_{\mu\nu\sigma\tau}$ must be a tensor density of weight $+1$. We can also see this by considering that $\epsilon_{\mu\nu\sigma\tau}$ must obey the covariant version of the transformation law (A.14)

$$\epsilon_{\mu'\nu'\sigma'\tau'} = [\det(\Lambda^{\mu}{}_{\mu'})]^{-1} \Lambda^{\mu}{}_{\mu'} \Lambda^{\nu}{}_{\nu'} \Lambda^{\sigma}{}_{\sigma'} \Lambda^{\tau}{}_{\tau'} \epsilon_{\mu\nu\sigma\tau}. \tag{A.17}$$

The matrix $\Lambda^{\mu}{}_{\mu'}$ is the inverse of $\Lambda^{\mu'}{}_\mu$, so $[\det(\Lambda^{\mu}{}_{\mu'})]^{-1} = [\det(\Lambda^{\mu'}{}_\mu)]^{+1}$. Given these weights, we see that the tensors $|g|^{-1/2}\epsilon^{\mu\nu\sigma\tau}$ and $|g|^{1/2}\epsilon_{\mu\nu\sigma\tau}$, which might be used to define dual tensors as in (3.82) and (3.83), transform without any factors of $\det(\Lambda)$.

A.3 Vector Spaces and Hilbert Spaces

Defined in an abstract way, a *linear vector space* is a collection of objects called *vectors*, for which I shall use the Dirac notation $|\ \rangle$, together with rules which allow two operations to be performed on them. The first operation is called *addition*: any two vectors $|a\rangle$ and $|b\rangle$ can be added to form a third vector, $|c\rangle = |a\rangle + |b\rangle$, which also belongs to the space. Just what this operation of addition means may depend on what the objects are that we want to identify as vectors. In the abstract, we require this operation to have the following four properties:

(i) addition is *commutative*, which means that $|a\rangle + |b\rangle = |b\rangle + |a\rangle$.
(ii) addition is *associative*, which means $(|a\rangle + |b\rangle) + |c\rangle = |a\rangle + (|b\rangle + |c\rangle)$.
(iii) the space contains a zero vector. I denote this by 0, without the $|\ \rangle$ symbol, because $|0\rangle$ is used in quantum theory for the quite different notion of a ground state or vacuum state. In the present context, nevertheless, 0 is a vector in the space. It has the property that $|a\rangle + 0 = |a\rangle$ for any vector $|a\rangle$.
(iv) given any vector $|a\rangle$ in the space, there also exists a unique vector $|-a\rangle$ such that $|a\rangle + |-a\rangle = 0$.

The second operation is *multiplication by scalars*. The scalars may be real numbers, in which case we have a *real vector space*, or complex numbers, in which case we have a *complex vector space*. Again, the exact effect of this multiplication may depend on what the vectors are, but in the abstract this operation is also required to have four properties:

(i) multiplication is *distributive* with respect to vectors, which means that $\alpha(|a\rangle + |b\rangle) = \alpha|a\rangle + \alpha|b\rangle$ for any two vectors $|a\rangle$ and $|b\rangle$ and any scalar α.
(ii) multiplication is also distributive with respect to scalars, which means that $(\alpha + \beta)|a\rangle = \alpha|a\rangle + \beta|a\rangle$ for any vector $|a\rangle$ and any two scalars α and β.
(iii) multiplication is associative, so that $\alpha(\beta|a\rangle) = (\alpha\beta)|a\rangle$.
(iv) multiplication by 1 leaves a vector unchanged, so $1|a\rangle = |a\rangle$.

Three-dimensional Euclidean space (or, for that matter, a d-dimensional Euclidean space) can be regarded as a real vector space if we choose one of its points as a preferred origin. Using Cartesian coordinates, the point with coordinates (x, y, z) corresponds to a vector $\boldsymbol{x}$, which can be conceived of as an arrow stretching from the origin to the point in question. It is easy to verify that the familiar parallelogram rule for adding vectors (by adding their components) and the rule for multiplying by a real number (which multiplies the length of the vector by that number, leaving its direction unchanged) satisfy the conditions listed above.

A vector space may, in addition, be equipped with a scalar product such as we introduced in §5.1. In the mathematical literature, it is more often called an *inner product* and the vector space is then an *inner product space*. As well as having the property (5.8), the inner product is required to be linear, in the sense that $(\alpha\langle a| + \beta\langle b|)|c\rangle = \alpha\langle a|c\rangle + \beta\langle b|c\rangle$. In Euclidean space, the usual 'dot product' of vectors, $\boldsymbol{x}\cdot\boldsymbol{x}' = xx' + yy' + zz'$ is a suitable inner product. An inner product can be thought of as conferring a *metric* on the space, which gives a notion of distance between two points,

$$d(a, b) = \sqrt{\langle a - b|a - b\rangle} \qquad \text{(A.18)}$$

where $|a - b\rangle$ means the vector $|a\rangle + |-b\rangle$. In the Euclidean vector space, for example, the quantity

$$d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(\boldsymbol{x} - \boldsymbol{x}') \cdot (\boldsymbol{x} - \boldsymbol{x}')} = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2} \tag{A.19}$$

is recognizable as the distance between the points $\boldsymbol{x}$ and $\boldsymbol{x}'$. Note, however, that this notion of distance is not quite the same as that defined by a metric tensor field in a manifold, which serves to define the length of a curve.

A *Hilbert space* can be defined as a complex vector space which possesses an inner product, but it is also required to be *complete*. Roughly speaking, completeness means that there are no vectors 'missing' from the space. More precisely, it is defined like this. We consider an infinite sequence of vectors $|a_n\rangle$ with the following property: given any real number ϵ, no matter how small, the distance $d(a_m, a_n)$ is less than ϵ, whenever m and n are greater than some value n_0, which may depend on ϵ. It is called a *Cauchy sequence*. A few moments thought should suggest the possibility that this sequence *converges* to a limiting vector $|a_\infty\rangle$, which would mean that $d(a_n, a_\infty) < \epsilon$, when $n > n_0$. The point is that this vector, towards which the sequence is 'trying' to converge might not exist, and this is what I mean by a 'missing' vector. Completeness means, then, that any Cauchy sequence of vectors actually does converge to a vector belonging to the Hilbert space.

A.4 Gauss' Theorem

The partial derivative of the metric determinant (A.15) is given by

$$\partial_\lambda g = \frac{1}{3!}\epsilon^{\mu\nu\sigma\tau}\epsilon^{\alpha\beta\gamma\delta}g_{\mu\alpha}g_{\nu\beta}g_{\sigma\gamma}(\partial_\lambda g_{\tau\delta}) = g g^{\mu\nu}(\partial_\lambda g_{\mu\nu}). \tag{A.20}$$

To see why the last expression is valid, consider that the quantity

$$e^{\tau\delta} = \frac{1}{3!}\epsilon^{\mu\nu\sigma\tau}\epsilon^{\alpha\beta\gamma\delta}g_{\mu\alpha}g_{\nu\beta}g_{\sigma\gamma}$$

is a symmetric rank $\binom{2}{0}$ tensor constructed from the metric, and must be proportional to $g^{\tau\delta}$. But $e^{\tau\delta}g_{\tau\delta}$ is equal to $4g$ and $g^{\tau\delta}g_{\tau\delta}$ is equal to 4, so the coefficient of proportionality is just g. Using the metric connection (2.50), we can calculate

$$\partial_\lambda(-g)^{1/2} = \tfrac{1}{2}(-g)^{1/2}g^{\mu\nu}(\partial_\lambda g_{\mu\nu}) = (-g)^{1/2}\Gamma^\mu{}_{\mu\lambda} \tag{A.21}$$

and armed with this result we can express the covariant divergence of a vector field as

$$\nabla_\mu V^\mu \equiv V^\mu{}_{;\mu} = \frac{1}{(-g)^{1/2}}\,\partial_\mu[(-g)^{1/2}V^\mu]. \tag{A.22}$$

The integral of the divergence of a vector field over a region D is a scalar quantity, provided that we use the covariant volume element (4.12). By using the version of Gauss' theorem which applies in Euclidean space, we can write it as a surface integral

$$\int_{\mathrm{D}} \mathrm{d}^4x\, (-g)^{1/2} V^{\mu}_{\ ;\mu} = \int_{\mathrm{D}} \mathrm{d}^4x\, \partial_\mu \left((-g)^{1/2} V^\mu \right) = \int_{\mathrm{S}} (-g)^{1/2} V^\mu \mathrm{d}S_\mu \qquad \text{(A.23)}$$

where S is the surface which bounds the region D.

A.5 Surface Area and Volume of a d-Dimensional Sphere

Let Ω_d be the surface area of a sphere of unit radius in d Euclidean dimensions. The surface area of a sphere of radius r is $\Omega_d r^{d-1}$ and we find by integrating this that its volume is $\Omega_d r^d/d$. To evaluate Ω_d, let $r^2 = x_1^2 + \ldots + x_d^2$ and consider the integral

$$\int_{-\infty}^{\infty} \mathrm{d}^d x\, \mathrm{e}^{-r^2} = \left[\int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-x^2} \right]^d = \pi^{d/2}. \qquad \text{(A.24)}$$

The solid angle subtended by the surface of the sphere at its centre is Ω_d, so if we change to polar coordinates and integrate over the $d - 1$ angular variables which do not appear in the integrand, this integral is

$$\pi^{d/2} = \Omega_d \int_0^\infty \mathrm{d}r\, r^{d-1} \mathrm{e}^{-r^2} = \tfrac{1}{2}\Omega_d \int_0^\infty \mathrm{d}t\, t^{d/2-1} \mathrm{e}^{-t} = \tfrac{1}{2}\Omega_d \Gamma(d/2) \qquad \text{(A.25)}$$

where $\Gamma(p) = (p - 1)!$ is Euler's gamma function. Thus, we have

$$\Omega_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}. \qquad \text{(A.26)}$$

Since $\Gamma(\frac{1}{2}) = \pi^{1/2}$ and $\Gamma(p + 1) = p\Gamma(p)$, we find, for example, that $\Omega_2 = 2\pi$, which is the circumference of a unit circle, and $\Omega_3 = 4\pi$, which is the surface area of a unit sphere in three dimensions. When carrying out spacetime integrals, we need to know that $\Omega_4 = 2\pi^2$.

A.6 Gaussian Integrals

Both in statistical mechanics and in quantum field theory, it is sometimes necessary to evaluate *Gaussian integrals*, the simplest example of which is

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-x^2} = \pi^{1/2}. \qquad \text{(A.27)}$$

A useful generalization is the integral

$$G(A) = \int \prod_{i=1}^{n} \mathrm{d}\phi_i^* \phi_i \, \exp\left(- \sum_{i,j=1}^{n} \phi_i^* A_{ij} \phi_j \right) \qquad \text{(A.28)}$$

where the ϕ_i are n complex variables and $\boldsymbol{A}$ is an Hermitian $n \times n$ matrix. To give this a precise meaning, let us say that $\phi_i = (\phi_{1i} + \mathrm{i}\phi_{2i})/\sqrt{2}$, where ϕ_{1i} and ϕ_{2i} are real variables, each to be integrated from $-\infty$ to ∞, and that the integration measure is $\mathrm{d}\phi_i^*\mathrm{d}\phi_i = \frac{1}{2}\mathrm{d}\phi_{1i}\mathrm{d}\phi_{2i}$. Other conventions may lead to integrals which differ from this one by a numerical factor; usually, this is not important because we are interested only in ratios of two such integrals or in derivatives of the logarithm of the integral as in (10.15).

To evaluate this integral, let us write the integrand as $\exp(-\boldsymbol{\phi}^\dagger \boldsymbol{A}\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is the column matrix whose elements are the ϕ_i. The matrix $\boldsymbol{A}$ can be diagonalized by the change of variable $\boldsymbol{\phi} = \boldsymbol{U}\boldsymbol{\psi}$, where $\boldsymbol{U}$ is a unitary matrix. Thus we have $\boldsymbol{\phi}^\dagger \boldsymbol{A}\boldsymbol{\phi} = \boldsymbol{\psi}^\dagger \boldsymbol{A}_{\mathrm{D}}\boldsymbol{\psi}$, where $\boldsymbol{A}_{\mathrm{D}}$ is a diagonal matrix, say with eigenvalues a_i. The Jacobian of this transformation is $\det(\boldsymbol{U}^\dagger\boldsymbol{U}) = \det(\boldsymbol{U}^{-1}\boldsymbol{U}) = 1$ and the integral becomes

$$G(\boldsymbol{A}) = (\tfrac{1}{2})^n \int_{-\infty}^{\infty} \prod_i \mathrm{d}\psi_{1i}\mathrm{d}\psi_{2i} \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n} a_i \left(\psi_{1i}^2 + \psi_{2i}^2\right)\right]. \qquad \text{(A.29)}$$

The further change of variables $\psi_{i1} \to (2/a_i)^{1/2}\psi_{i1}$ and $\psi_{i2} \to (2/a_i)^{1/2}\psi_{i2}$ converts this expression to a product of $2n$ integrals of the form (A.27):

$$G(\boldsymbol{A}) = \prod_{i=1}^{n} \left[a_i^{-1/2} \int \mathrm{d}\psi_{1i}\, \mathrm{e}^{-\psi_{1i}^2}\right]\left[a_i^{-1/2} \int \mathrm{d}\psi_{2i}\, \mathrm{e}^{-\psi_{2i}^2}\right] = \frac{\pi^n}{\det(\boldsymbol{A})} \qquad \text{(A.30)}$$

because $\det(\boldsymbol{A}) = \prod_i a_i$. With care, a functional integral such as (10.80), where $\boldsymbol{A}$ is a differential operator rather than a matrix, can be evaluated in a similar way, although we have seen in chapters 9 and 10 that the explicit evaluation of these integrals can often be avoided.

A.7 Grassmann Variables

A set of variables θ_i which anticommute with each other, so that $\theta_i\theta_j = -\theta_j\theta_i$ is said to generate a *Grassmann algebra.* This algebra is a special kind of linear vector space (see appendix A.3) so it is implied that Grassmann numbers can be added and multiplied by scalars, which might be either real or complex numbers. In the applications that concern us, there are, say, $2n$ of these variables, which might be regarded as two sets of real variables, b_i and c_i, with $i = 1, \ldots, n$ or as n complex variables θ_i and their complex conjugates $\bar{\theta}_i$. To be definite, I take the latter view.

As with matrices and operators, functions of Grassmann variables are defined by means of Taylor series. The square of any Grassmann variable must be zero, because $\theta_i\theta_i = -\theta_i\theta_i$, so if n is finite, any Taylor series has only a finite number of terms—those in which each variable appears at most once. For example, any function of a single Grassmann variable and its complex conjugate can be written as

$$f(\theta,\bar{\theta}) = f_0 + f_1\theta + f_2\bar{\theta} + f_3\bar{\theta}\theta \tag{A.31}$$

where $f_0, \dots, f_3$ are complex numbers. Consider, in particular, the function

$$\exp\left(\sum_{i,j=1}^{n} \bar{\theta}_i A_{ij}\theta_j\right) = 1 + \ldots + \mathcal{A}(\bar{\theta}_1\theta_1)\cdots(\bar{\theta}_n\theta_n) \tag{A.32}$$

where $\boldsymbol{A}$ is an $n \times n$ matrix. For $n = 2$, it is simple to calculate the coefficient $\mathcal{A}$ of the last term. We have

$$\begin{aligned}\mathcal{A}(\bar{\theta}_1\theta_1)(\bar{\theta}_2\theta_2) &= \tfrac{1}{2}\left[A_{11}(\bar{\theta}_1\theta_1) + A_{12}(\bar{\theta}_1\theta_2) + A_{21}(\bar{\theta}_2\theta_1) + A_{22}(\bar{\theta}_2\theta_2)\right]^2 \\ &= \left[A_{11}A_{22}(\bar{\theta}_1\theta_1)(\bar{\theta}_2\theta_2) + A_{12}A_{21}(\bar{\theta}_1\theta_2)(\bar{\theta}_2\theta_1)\right] \\ &= \left[A_{11}A_{22} - A_{12}A_{21}\right](\bar{\theta}_1\theta_1)(\bar{\theta}_2\theta_2)\end{aligned} \tag{A.33}$$

and so $\mathcal{A} = \det(\boldsymbol{A})$. Readers should not find it hard to convince themselves that this result is valid for any n.

Differentiation with respect to a Grassmann variable can be defined in the following way. Any function depending on all the θ_i can be decomposed as $f(\{\theta_i\}) = f_0(\{\theta_{i\neq j}\}) + \theta_j f_1(\{\theta_{i\neq j}\})$, where f_0 and f_1 are independent of the particular variable θ_j with respect to which we want to differentiate. Then we will say that

$$\frac{\partial}{\partial\theta_j} f(\{\theta_i\}) = f_1(\{\theta_{i\neq j}\}). \tag{A.34}$$

Note carefully that θ_j stands to the left of all the other Grassmann variables contained in f_1. Given an expression for f in which this is not true, we must move θ_j to the leftmost position, taking account of all the $-$ signs that arise from anticommutations. Using this rule, it is a simple exercise to show that partial derivatives anticommute:

$$\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} = -\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}. \tag{A.35}$$

The definition of integration with respect to a Grassmann variable is easily stated: the symbol $\int \mathrm{d}\theta_i$ means *exactly the same* as $\partial/\partial\theta_i$. The temptation to wonder whether this is a ‘correct’ generalization of the usual notion of integration with respect to an ordinary variable is one that should be resisted. The plain fact is that, by adopting this definition, we arrive at a path integral of the form

$$\int \mathcal{D}\psi\mathcal{D}\bar{\psi}\ \psi(x_1)\cdots\bar{\psi}(x_m)\exp[\mathrm{i}S(\psi,\bar{\psi})] \tag{A.36}$$

which correctly represents fermionic matrix elements of the form $\langle 0|\psi(x_1)\cdots\bar{\psi}(x_m)|0\rangle$, and it is for this reason that the definition is a useful one. (A convincing proof of this plain fact is quite involved, and I shall not attempt one here. A transparent discussion is not easy to find in the literature, but interested readers may like to consult Itzykson and Zuber (1980).) To the extent that a Grassmann integral can be thought of by analogy with an ordinary integral, it is the counterpart of a definite integral $\int_{-\infty}^{\infty} \mathrm{d}x$; the notion of an indefinite integral, which is the inverse operation to differentiation has no useful generalization to Grassmann variables.

Most often, as in (A.36), we want to integrate over all of the variables in the Grassmann algebra. When there are finitely many of these, the answer is just the last term in the Taylor series for the function we want to integrate. Specifically, we have

$$\int (\mathrm{d}\theta_n \mathrm{d}\bar{\theta}_n) \cdots (\mathrm{d}\theta_1 \mathrm{d}\bar{\theta}_1)\, (\bar{\theta}_1 \theta_1) \cdots (\bar{\theta}_n \theta_n) = 1 \tag{A.37}$$

and the integrals of terms from which one or more of the θ_i or $\bar{\theta}_i$ are missing give zero. In fact, the only integral for which we ordinarily need an explicit result is the integral of a Gaussian function such as (A.32), for which the answer is $\det(\boldsymbol{A})$. Up to a factor of π^n, which can generally be absorbed into a normalizing constant, this is the inverse of the corresponding ‘bosonic’ integral (A.28). Integrals such as (15.136) and (15.137) are related in the same way, once we give them a definite meaning through a Wick rotation to Euclidean space (see §15.2.3).

Appendix B

Some Elements of Group Theory

The mathematical framework which allows a systematic study of the consequences of symmetry in physics is *group theory*. In the main text, I have drawn on various group-theoretical ideas in an *ad hoc* way, as the occasion demanded. In this appendix, I attempt to draw together some of the essential features of group theory in a more coherent way, relying largely on the example of rotations in three dimensions. I do not, however, have the space to develop in detail the extensive body of techniques and results to which these ideas give rise. Readers who would like to know more will find group theory discussed at varied levels of sophistication both in specialized books devoted to that topic and in more summary form in books on particle physics and quantum field theory. A small selection of useful sources is: Cheng and Li (1984), Coleman (1985), Cornwell (1984), de Azcárraga and Izquierdo (1995), Halzen and Martin (1984), Jones (1998), Nakahara (1990), Tung (1985), Ticciati (1999).

Abstractly defined, a *group G* is a collection of elements g with the following properties:

(i) there is a rule for multiplying any two elements and their product g_1g_2 is also an element of G. This rule for multiplication is *associative*, so for any three elements $(g_1g_2)g_3 = g_1(g_2g_3)$.
(ii) there is an *identity element* of G, say e, such that $eg = ge = g$ for any element g.
(iii) for every element g, there is a unique inverse element g^{-1} such that $gg^{-1} = g^{-1}g = e$.

The rule for multiplication is not necessarily commutative. That is, g_1g_2 is not necessarily the same as g_2g_1. If $g_1g_2 = g_2g_1$ for every pair of elements, then the group is said to be *Abelian*; otherwise it is *non-Abelian*.

In most applications to physics, the elements of a group are transformations of some kind. Very often, a group of transformations has infinitely many elements, labelled by one or more parameters which can assume a continuous range of values, such as the three components of a vector $\boldsymbol{a}$ that specify a spatial translation $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{a}$ or the angles that specify a rotation. A group of this kind is called a *Lie group*. The multiplication of elements corresponds to successive applications of two transformations. For example, if we write the operation of spatial translation of a position vector as $g(\boldsymbol{a})\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{a}$, then a sequence of two translations gives $g(\boldsymbol{b})g(\boldsymbol{a})\boldsymbol{x} = g(\boldsymbol{b})(\boldsymbol{x} + \boldsymbol{a}) = \boldsymbol{x} + \boldsymbol{a} + \boldsymbol{b}$. The net effect is a translation through the vector $\boldsymbol{a} + \boldsymbol{b}$, so property (i) is satisfied. In fact, we have $g(\boldsymbol{b})g(\boldsymbol{a}) = g(\boldsymbol{a} + \boldsymbol{b}) = g(\boldsymbol{a})g(\boldsymbol{b})$, so the group of space translations is Abelian. Properties (ii) and (iii) are also satisfied: the identity element $e = g(\boldsymbol{0})$, namely a translation through a vector of zero length, leaves any vector $\boldsymbol{x}$ unchanged, and we can obviously identify the inverse elements as $g^{-1}(\boldsymbol{a}) = g(-\boldsymbol{a})$.

Let us now focus on the less trivial example of rotations. We shall regard Euclidean space as a vector space, as in appendix A.3, with a fixed origin for Cartesian coordinates, which I will call x^1, x^2 and x^3. Let $\boldsymbol{r}$ be a position vector, with components (x^1, x^2, x^3). A rotation about the x^3 axis through an angle α leads to a new set of components

$$x^{1'} = x^1 \cos\alpha + x^2 \sin\alpha \qquad x^{2'} = -x^1 \sin\alpha + x^2 \cos\alpha \qquad x^{3'} = x^3. \quad \text{(B.1)}$$

If we represent $\boldsymbol{r}$ as a column matrix, this can be written as

$$\boldsymbol{r}' = R_3(\alpha)\boldsymbol{r} \qquad \text{(B.2)}$$

with

$$\boldsymbol{r} = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} \qquad \boldsymbol{r}' = \begin{pmatrix} x^{1'} \\ x^{2'} \\ x^{3'} \end{pmatrix} \qquad R_3(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad \text{(B.3)}$$

The new column matrix $\boldsymbol{r}'$ can be regarded either as giving the components of a new vector, obtained by rotating $\boldsymbol{r}$ through an angle $-\alpha$ (the *active* point of view) or as giving the components of the same vector relative to a new set of axes, obtained by rotating the old axes through an angle $+\alpha$ (the *passive* point of view).

It is often helpful to consider a rotation through a finite angle to be made up of a sequence of infinitesimal rotations. If the angle α is infinitesimal, then we can write

$$R_3(\alpha) = I + \mathrm{i}\alpha\mathcal{J}_3 + \mathrm{O}(\alpha^2) \qquad \text{(B.4)}$$

where the matrix

$$\mathcal{J}^3 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{(B.5)}$$

is called the *generator* of rotations about the x^3 axis and I is the unit 3×3 matrix. To build up a rotation through a finite angle α, we can rotate N times through the angle α/N, which is very small when N is very large. The identity

$$\lim_{N\to\infty} \left(I + \mathrm{i}\frac{\alpha}{N}\mathcal{J}^3\right)^N = \exp(\mathrm{i}\alpha\mathcal{J}^3) \qquad \text{(B.6)}$$

shows that $R_3(\alpha)$ ought to be equal to $\exp(\mathrm{i}\alpha\mathcal{J}^3)$. Readers should find it instructive to verify this explicitly by working out the matrix $(\mathcal{J}^3)^2$ and verifying that the Taylor series

$$I+\sum_{n=1}^{\infty}\frac{1}{n!}(\mathrm{i}\alpha)^n(\mathcal{J}^3)^n = I+(\cos\alpha-1)(\mathcal{J}^3)^2+\mathrm{i}\sin\alpha\,\mathcal{J}^3 \tag{B.7}$$

does indeed reproduce the matrix $R_3(\alpha)$.

For rotations about the x^1 and x^2 axes, the analogous generator matrices are

$$\mathcal{J}^1=\begin{pmatrix}0&0&0\\0&0&-\mathrm{i}\\0&\mathrm{i}&0\end{pmatrix}\qquad \mathcal{J}^2=\begin{pmatrix}0&0&\mathrm{i}\\0&0&0\\-\mathrm{i}&0&0\end{pmatrix}. \tag{B.8}$$

More generally, we can consider a rotation through an angle α about an axis in the direction of a unit vector $\boldsymbol{n}$. The rotation matrix that does this can be written down by defining a vector of three angles $\boldsymbol{\alpha}=(\alpha^1,\alpha^2,\alpha^3)$, such that $\boldsymbol{\alpha}=\alpha\boldsymbol{n}$ and the vector of generator matrices $\boldsymbol{\mathcal{J}}=(\mathcal{J}^1,\mathcal{J}^2,\mathcal{J}^3)$. Then the desired matrix is

$$R(\boldsymbol{\alpha})=\exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}}). \tag{B.9}$$

Intuitively, it is fairly obvious that the net effect of two successive rotations, possibly through different angles and about different axes, is a rotation through some angle about some axis. It is necessary that this should be so if the collection of all rotations is to form a group (in particular, if property (i) is to be satisfied). Thus, given two vector angles $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, there must exist a third one, $\boldsymbol{\gamma}$, such that

$$\exp(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}})\exp(\mathrm{i}\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}})=\exp(\mathrm{i}\boldsymbol{\gamma}\cdot\boldsymbol{\mathcal{J}}). \tag{B.10}$$

If this is to be true, then the matrices $\mathcal{J}^i$ must have a certain property, which can be found as follows. Treating the angles α^i and β^i as small, the left-hand side of (B.10) can be expanded as a Taylor series, whose first few terms are

$$I+\mathrm{i}(\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}}+\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}})-\tfrac{1}{2}\left[(\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}})^2+2(\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}})(\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}})+(\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}})^2\right]+\ldots. \tag{B.11}$$

The logarithm of this quantity can be found using the expansion $\ln(I+X)=X-\frac{1}{2}X^2+\ldots$ and is given by

$$\mathrm{i}(\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}}+\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}})-\tfrac{1}{2}[\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}},\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{J}}]+\ldots=\mathrm{i}(\alpha^i+\beta^i)\mathcal{J}^i-\tfrac{1}{2}\alpha^i\beta^j[\mathcal{J}^i,\mathcal{J}^j]+\ldots. \tag{B.12}$$

If this is to be expressible as $\mathrm{i}\boldsymbol{\gamma}\cdot\boldsymbol{\mathcal{J}}$, which is a linear combination of the $\mathcal{J}^i$, then the commutator $[\mathcal{J}^i,\mathcal{J}^j]$ must be a linear combination of the $\mathcal{J}^i$, say

$$[\mathcal{J}^i,\mathcal{J}^j]=\mathrm{i}C^{ijk}\mathcal{J}^k. \tag{B.13}$$

If this is true, then it turns out that all the remaining terms can also be written as a linear combination of the $\mathcal{J}^i$. This argument does not depend on the $\mathcal{J}^i$

being the generators of rotations, so the generators of *any* Lie group must have commutation relations of this kind. The coefficients C^{ijk} are called the *structure constants* and their numerical values largely determine the properties of the Lie group in question. Because the left-hand side of (B.13) is a commutator, it is clear that $C^{ijk} = -C^{jik}$, and it can be shown that (with a suitable choice of the generators) C^{ijk} is in fact totally antisymmetric. In the case of rotations, the commutators are easily worked out from (B.5) and (B.8). We find

$$[\mathcal{J}^1, \mathcal{J}^2] = \mathrm{i}\mathcal{J}^3 \qquad [\mathcal{J}^2, \mathcal{J}^3] = \mathrm{i}\mathcal{J}^1 \qquad [\mathcal{J}^3, \mathcal{J}^1] = \mathrm{i}\mathcal{J}^2 \tag{B.14}$$

which shows that the structure constants of the rotation group are $C^{ijk} = \epsilon^{ijk}$. For any Lie group, say with generators $\mathcal{T}^a$, the collection of all possible linear combinations of generators $\alpha^a\mathcal{T}^a$ constitutes what is called the *Lie algebra*. (An *algebra* is a vector space, in the general sense discussed in appendix A.3, which possesses an additional structure, here represented by the fact that the vectors are matrices with the commutation relations (B.13).)

Objects other than vectors will transform in other ways under rotations. Consider, for example, a rank-2 tensor T^{ij}. Adapting the general transformation law (2.19) to our present notation, we see that

$$T^{i'j'} = R^{i'i}(\boldsymbol{\alpha})R^{j'j}(\boldsymbol{\alpha})T^{ij}. \tag{B.15}$$

If we regard the components T^{ij} as the elements of a 3×3 matrix T, then this can be written as

$$T' = R(\boldsymbol{\alpha})TR^{\mathrm{T}}(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha})TR^{-1}(\boldsymbol{\alpha}). \tag{B.16}$$

I have used the fact that R is an orthogonal matrix so that $R^{\mathrm{T}} = R^{-1}$; this is easily verified for the particular matrix given in (B.3) and I shall discuss the general case a little further below. On the other hand, we might assemble the nine components T^{ij} into a column matrix, say $\boldsymbol{T}$. The new column matrix $\boldsymbol{T}'$ resulting from a rotation must be expressible as

$$\boldsymbol{T}' = R^{(9)}(\boldsymbol{\alpha})\boldsymbol{T} \qquad R^{(9)}(\boldsymbol{\alpha}) = \exp\left(\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\mathcal{J}}^{(9)}\right) \tag{B.17}$$

where $R^{(9)}$ is a 9×9 matrix, and $\boldsymbol{\mathcal{J}}^{(9)}$ denotes a set of three 9×9 matrices which obey the same commutation relations (B.14) as the original $\boldsymbol{\mathcal{J}}$.

In general, any set of matrices having the commutation relations (B.13) appropriate to a particular Lie group is said to constitute a *representation* of the Lie algebra and these matrices generate the transformations of some kind of tensor. In much of the physics literature, the tensor which transforms using a particular representation of the generators is also referred to as the representation. In many instances, two representations differ from each other in so trivial a way that they may be regarded as equivalent. Consider, for example the column matrix

$$\bar{\boldsymbol{r}} = S\boldsymbol{r} = \begin{pmatrix} (x^1 - \mathrm{i}x^2)/\sqrt{2} \\ x^3 \\ -(x^1 + \mathrm{i}x^2)/\sqrt{2} \end{pmatrix} \qquad S = \begin{pmatrix} 1/\sqrt{2} & -\mathrm{i}/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & -\mathrm{i}/\sqrt{2} & 0 \end{pmatrix}. \tag{B.18}$$

My reason for choosing this particular matrix S will become apparent later, but it is clear that $\boldsymbol{r}$ and $\bar{\boldsymbol{r}}$ contain exactly the same information: they differ only in the way this information is distributed between the elements of the matrix. Under a rotation, we have

$$\bar{\boldsymbol{r}}' = S\boldsymbol{r}' = SR(\boldsymbol{\alpha})\boldsymbol{r} = SR(\boldsymbol{\alpha})S^{-1}\bar{\boldsymbol{r}} = \bar{R}(\boldsymbol{\alpha})\bar{\boldsymbol{r}} \tag{B.19}$$

where $\bar{R}(\boldsymbol{\alpha}) = SR(\boldsymbol{\alpha})S^{-1}$. In this case, the inverse matrix is

$$S^{-1} = \begin{pmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ \mathrm{i}/\sqrt{2} & 0 & \mathrm{i}/\sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}. \tag{B.20}$$

A little thought will show that the new rotation matrix can be written as $\bar{R}(\boldsymbol{\alpha}) = \exp(\mathrm{i}\boldsymbol{\alpha}\cdot\bar{\boldsymbol{\mathcal{J}}})$ with generator matrices

$$\bar{\mathcal{J}}^i = S\mathcal{J}^i S^{-1}. \tag{B.21}$$

The two sets of generators are said to be related by a *similarity transformation* and two representations which are related by a similarity transformation (where S may be any matrix whose inverse is well defined) are said to be *equivalent*. It is not hard to see that if $\mathcal{J}^i$ are any matrices with the commutation relations (B.13), then the matrices $S\mathcal{J}^i S^{-1}$ have the same commutation relations.

Rotations also affect functions of the coordinates, such as the wavefunctions for quantum-mechanical particles. From the passive point of view, a scalar function $\psi(\boldsymbol{r})$ will be expressed in a rotated coordinate system by a new function $\psi'(\boldsymbol{r}')$, such that $\psi'(\boldsymbol{r}') = \psi(\boldsymbol{r})$ when $\boldsymbol{r}$ and $\boldsymbol{r}'$ are the old and new coordinates of the same point. If we rewrite this as $\psi'(\boldsymbol{r}') = \psi(R^{-1}\boldsymbol{r}')$, then $\boldsymbol{r}'$ is just a dummy variable, and we can drop the prime. If α is infinitesimal in (B.1), then we can use a Taylor series to express the transformation as

$$\begin{aligned}\psi'(x^1, x^2, x^3) &= \psi(x^1 - \alpha x^2, x^2 + \alpha x^1, x^3) \\ &= [1 + \mathrm{i}\alpha\mathcal{J}^3]\psi(x^1, x^2, x^3)\end{aligned} \tag{B.22}$$

where the generator is now the differential operator

$$\mathcal{J}^3 = -\mathrm{i}\left(x^1\frac{\partial}{\partial x^2} - x^2\frac{\partial}{\partial x^1}\right). \tag{B.23}$$

This, together with the analogous generators of rotations about the x^2 and x^3 axes can be summarized as $\mathcal{J}^i = -\mathrm{i}\epsilon^{ijk}x^j\partial/\partial x^k$. We should not be surprised to find that they satisfy the commutation relations (B.14) and thus furnish a representation of the Lie algebra.

Some degree of order can be imposed upon the vast collection of possible representations of a given group by the idea of an *irreducible* representation.

Consider, for example, a rank-2 tensor whose transformation law is (B.16) or (B.17). Its components can be split into three sets by defining

$$T_0 = T^{ii} \qquad T_{\mathrm{S}}^{ij} = \tfrac{1}{2}\left(T^{ij} + T^{ji}\right) - \tfrac{1}{3}T^{kk}\delta^{ij} \qquad T_{\mathrm{A}}^{ij} = \tfrac{1}{2}\left(T^{ij} - T^{ji}\right). \tag{B.24}$$

Of these three tensors, T_0, the trace of the matrix T, is a scalar, T_{S} is a symmetric, traceless tensor, with five independent components and T_{A} is an antisymmetric tensor with three independent components. Together, they account for the nine degrees of freedom in T^{ij}, but under a rotation they transform independently. That is to say, the antisymmetric part of $\boldsymbol{T}'$ is a rearrangement of the antisymmetric part of $\boldsymbol{T}$ and so on. In forming the nine-component column matrix $\boldsymbol{T}$, we can choose to list first T_0, then the components of T_{A} and finally those of T_{S}. Then one of the generators $\mathcal{J}^{(9)}$ will have the block-diagonal form

$$\mathcal{J}^{(9)} = \begin{pmatrix} \mathcal{J}^{(1)} & 0 & 0 \\ 0 & \mathcal{J}^{(3)} & 0 \\ 0 & 0 & \mathcal{J}^{(5)} \end{pmatrix}. \tag{B.25}$$

Since T_0 is a scalar, unchanged by the rotation, the three 1×1 matrices $\boldsymbol{\mathcal{J}}^{(1)}$ are equal to 0. They satisfy the commutation relations, but in a trivial way, and constitute what is called an 'unfaithful' representation. Neither the 3×3 matrices $\boldsymbol{\mathcal{J}}^{(3)}$ nor the 5×5 matrices $\boldsymbol{\mathcal{J}}^{(5)}$ can be further decomposed in the same way. They are said to constitute *irreducible* representations of the rotation group, while the $\boldsymbol{\mathcal{J}}^{(9)}$ constitute a reducible representation. In particular, the tensor T might be the 'direct product' of two vectors, say $\boldsymbol{u}$ and $\boldsymbol{v}$, which means that its components are $T^{ij} = u^i v^j$. The fact that it can be decomposed into irreducible tensors as in (B.24) might then be expressed as

$$\mathbf{3} \otimes \mathbf{3} = \mathbf{1} \oplus \mathbf{3} \oplus \mathbf{5} \tag{B.26}$$

though many variants of this kind of symbolism are to be found in the literature. Systematic methods for obtaining such decompositions are described in the books mentioned above.

In the case of rotations, it is fairly clear that the three generators can be assembled into a vector of matrices, for which I have used the suggestive boldface notation $\boldsymbol{\mathcal{J}}$, which ought to transform in the same way as the position vector $\boldsymbol{r}$ under rotations. To make this explicit, consider a passive rotation of the coordinate axes through an angle $\boldsymbol{\beta}$, so that a position vector has components $\boldsymbol{r}$ relative to the old axes and $\widetilde{\boldsymbol{r}}$ relative to the new ones, with $\widetilde{\boldsymbol{r}} = R(\boldsymbol{\beta})\boldsymbol{r}$. Let us rotate this vector through an angle, and about an axis, specified by a vector which has components $\boldsymbol{\alpha}$ relative to the old axes, and therefore has components $\widetilde{\boldsymbol{\alpha}} = R(\boldsymbol{\beta})\boldsymbol{\alpha}$ relative to the new ones. The rotated vector has components $R(\boldsymbol{\alpha})\boldsymbol{r}$ relative to the old axes and $R(\widetilde{\boldsymbol{\alpha}})\widetilde{\boldsymbol{r}}$ relative to the new ones, so we have

$$R(\widetilde{\boldsymbol{\alpha}})\widetilde{\boldsymbol{r}} = R(\boldsymbol{\beta})R(\boldsymbol{\alpha})\boldsymbol{r} = R(\boldsymbol{\beta})R(\boldsymbol{\alpha})R^{-1}(\boldsymbol{\beta})\widetilde{\boldsymbol{r}}. \tag{B.27}$$

This must be true for any vector $\widetilde{\boldsymbol{r}}$ so, taking $\boldsymbol{\alpha}$ to be infinitesimal, we find

$$\widetilde{\boldsymbol{\alpha}} \cdot \boldsymbol{\mathcal{J}} = R(\boldsymbol{\beta})\boldsymbol{\alpha} \cdot \boldsymbol{\mathcal{J}} R^{-1}(\boldsymbol{\beta}). \tag{B.28}$$

The components of $\widetilde{\boldsymbol{\alpha}}$ are $\widetilde{\alpha}^j = R^{ji}(\boldsymbol{\beta})\alpha^i$, so we get

$$R(\boldsymbol{\beta})\mathcal{J}^i R^{-1}(\boldsymbol{\beta}) = R^{ji}(\boldsymbol{\beta})\mathcal{J}^j = R^{ij}(-\boldsymbol{\beta})\mathcal{J}^j. \tag{B.29}$$

The expression on the right-hand side rearranges the generators $\mathcal{J}^i$ in the same way that a rotation rearranges the components of a vector. By taking $\boldsymbol{\beta}$ to be infinitesimal, we obtain the commutation relation

$$[\mathcal{J}^i, \mathcal{J}^j] = \left(\mathcal{J}^j\right)^{ik} \mathcal{J}^k \tag{B.30}$$

where $\left(\mathcal{J}^j\right)^{ik}$ means the ikth element of the matrix $\mathcal{J}^j$. To agree with (B.14), we must have $\left(\mathcal{J}^j\right)^{ik} = \mathrm{i}\epsilon^{ijk}$ and this does indeed reproduce the matrices (B.5) and (B.8). We see that there is a special representation, which transforms both vectors and the generators themselves, in which the generator matrices can be constructed from the structure constants. The same is true for any Lie group and the special representation is called the *adjoint* representation. It is discussed from a slightly different point of view in chapter 8 (see the discussion of (8.31) and exercise 8.4).

In quantum mechanics, the operators which represent the Cartesian components of angular momentum are defined in terms of the rotation generators as $\hat{J}^i = \hbar\hat{\mathcal{J}}^i$. This means that the $\hat{J}^i$ are operators in the Hilbert space of state vectors which satisfy the commutation relations

$$[\hat{J}^i, \hat{J}^j] = \mathrm{i}\hbar\epsilon^{ijk}\hat{J}^k. \tag{B.31}$$

The transformations of operators associated with other physical quantities are given by expressions similar to (B.29). For example, the state vectors representing states of a single particle with momenta $\boldsymbol{p}$ and $R(\boldsymbol{\alpha})\boldsymbol{p}$ are related by

$$\exp\left(\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right)|\boldsymbol{p}\rangle = |R(\boldsymbol{\alpha})\boldsymbol{p}\rangle. \tag{B.32}$$

Acting on each side with the momentum operator $\hat{\boldsymbol{p}}$, we find

$$\begin{aligned}\hat{p}^i \exp\left(\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right)|\boldsymbol{p}\rangle &= R(\boldsymbol{\alpha})^{ij} p^j |R(\boldsymbol{\alpha})\boldsymbol{p}\rangle \\ &= R(\boldsymbol{\alpha})^{ij} \exp\left(\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right) p^j |\boldsymbol{p}\rangle \\ &= R(\boldsymbol{\alpha})^{ij} \exp\left(\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right) \hat{p}^j |\boldsymbol{p}\rangle \end{aligned} \tag{B.33}$$

and thus

$$\exp\left(-\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right)\hat{p}^i \exp\left(\frac{\mathrm{i}}{\hbar}\boldsymbol{\alpha} \cdot \hat{\boldsymbol{J}}\right) = R^{ij}(\boldsymbol{\alpha})\hat{p}^j. \tag{B.34}$$

By taking $\boldsymbol{\alpha}$ to be infinitesimal, we find the commutation relation

$$[\hat{p}^i, \hat{J}^j] = \mathrm{i}\hbar\epsilon^{ijk}\hat{p}^k. \tag{B.35}$$

If the particle's angular momentum arises purely from its orbital motion, we should have

$$\hat{\boldsymbol{J}} = \hat{\boldsymbol{x}} \times \hat{\boldsymbol{p}} \qquad \text{or} \qquad \hat{J}^i = \epsilon^{ijk}\hat{x}^j\hat{p}^k \tag{B.36}$$

and readers may easily check that the commutators (B.31) and (B.35) are consistent with the basic canonical commutator (5.38). In the remainder of this appendix, I deal only with quantum operators, and will once more omit the circumflex.

Quite often, we need to know the eigenvalues and eigenstates of the angular momentum operators. These fall into multiplets corresponding to the irreducible representations of the rotation group. Similarly, the multiplets of particles encountered in theories with non-Abelian gauge symmetries correspond to the irreducible representations of the appropriate symmetry group. As we learned in chapter 5, only operators which commute with each other can have simultaneous eigenstates. According to (B.31), no two of the rotation generators commute with each other, so we look for eigenstates of just one of them. Conventionally, we use J^3, and the x^3 axis singled out in this way is sometimes referred to as the *spin quantization axis*. This axis is singled out only by the way in which we choose to describe a system, and has no physical meaning. On the other hand, if a special direction in space is singled out by physical circumstances, such as an external magnetic or electric field applied to the system of interest, then it is usually convenient to choose this direction as the quantization axis. Although no two of the J^i commute with each other, there is another operator, namely $\boldsymbol{J}^2 = (J^1)^2 + (J^2)^2 + (J^3)^2$, which commutes with all of them. In general, an operator which commutes with all the generators is called a *Casimir* operator (after H Casimir). The rotation group has only one Casimir operator, but other groups may have several. Each irreducible representation corresponds to a definite value of every Casimir operator.

The eigenvalues and eigenvectors of $\boldsymbol{J}^2$ and J^3 can be found by the same method that we used in chapter 5 to find the energy levels of the harmonic oscillator. The two operators

$$J^{\pm} = J^1 \pm \mathrm{i}J^2 \tag{B.37}$$

have the commutation relations

$$[J^+, J^-] = 2\hbar J^3 \tag{B.38}$$

$$[J^{\pm}, J^3] = \mp\hbar J^{\pm}. \tag{B.39}$$

Comparing (B.39) with (5.60) and (5.61) we see that J^+ acts as a raising operator for the eigenvalue of J^3, while J^- acts as a lowering operator. The operator

$\boldsymbol{J}^2$, representing the square of the total angular momentum, can be expressed by means of (B.38) as

$$\boldsymbol{J}^2 = J^+J^- + J^3(J^3 - \hbar) = J^-J^+ + J^3(J^3 + \hbar) \tag{B.40}$$

and for a given value of $\boldsymbol{J}^2$, there must be a maximum value of J^3, say $j\hbar$. We look, then, for states $|j, m\rangle$, such that $J^3|j, m\rangle = m\hbar|j, m\rangle$. The state $|j, j\rangle$, in which m has its maximum value of j, has the property $J^+|j, j\rangle = 0$ and the second expression in (B.40) shows that the eigenvalue of $\boldsymbol{J}^2$ is $j(j+1)\hbar^2$. Similarly, m has a minimum value, say $m_{\min}$, for which $J^-|j, m_{\min}\rangle = 0$, and we deduce from the first equality in (B.40) that $m_{\min} = -j$. We thus find multiplets, each corresponding to a definite value of j, within which m takes values ranging from $-j$ to j in integer steps. Each multiplet contains $(2j+1)$ states, so j must be either an integer or a half-odd-integer.

In the case of orbital angular momentum, the operators (B.36) can be realized as differential operators that act on wavefunctions. The operators J^i are

$$J^i = -\mathrm{i}\hbar\epsilon^{ijk}x^j\partial/\partial x^k \tag{B.41}$$

(see (B.23) and the following discussion). By solving the equations

$$J^3\psi_{jm}(\boldsymbol{x}) = m\hbar\psi_{jm}(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{J}^2\psi_{jm}(\boldsymbol{x}) = j(j+1)\hbar^2\psi_{jm}(\boldsymbol{x}) \tag{B.42}$$

one finds that only integer values of j and m are allowed. In the case $j = 1$, the eigenfunctions are given by

$$\psi_{11}(\boldsymbol{x}) \propto (x^1 + \mathrm{i}x^2) \qquad \psi_{10}(\boldsymbol{x}) \propto x^3 \qquad \psi_{1\,-1}(\boldsymbol{x}) \propto (x^1 - \mathrm{i}x^2). \tag{B.43}$$

These are more usually expressed in polar coordinates in terms of the *spherical harmonics* $Y_{jm}(\theta, \phi)$, but it is interesting to observe that these eigenfunctions are just the coordinates introduced in (B.18). Correspondingly, the generator matrices defined in (B.21) are given by

$$\bar{\mathcal{J}}^+ = -\sqrt{2}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad \bar{\mathcal{J}}^- = -\sqrt{2}\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$
$$\bar{\mathcal{J}}^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{B.44}$$

Evidently, using this complex coordinate basis, the generator $\bar{\mathcal{J}}^3$ is diagonal; its diagonal elements are the eigenvalues $m = 1, 0, -1$ and its eigenvectors $|j, m\rangle$ are

$$|1, 1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad |1, 0\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad |1, -1\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{B.45}$$

The action of $\bar{\mathcal{J}}^{\pm}$ on these eigenvectors is easily checked. It is apparent that both the position vector $\boldsymbol{r}$ and the set of eigenfunctions (B.43) are manifestations of the $j = 1$ representation of the rotation group, and I leave it as an exercise for interested readers to explore exactly how this equivalence works. The five eigenfunctions $\psi_{2m}(\boldsymbol{x})$ for $m = -2, \ldots, 2$ correspond in a similar way to the five independent components of the traceless, symmetric part of the tensor $T^{ij} = x^i x^j$ (see equation (B.24)).

In non-relativistic quantum mechanics, we do not know *a priori* whether the half-odd-integer values of j have any relevance to physics. At any rate, they cannot describe orbital angular momentum. As it turns out, they are relevant for describing the *intrinsic* angular momentum or *spin* of certain particles, the most familiar of which are electrons, protons and neutrons, for which (using s for spin in place of j) $s = \frac{1}{2}$. It is customary to describe spin in the non-relativistic theory by using a two-component wavefunction

$$\psi(\boldsymbol{x}) = \begin{pmatrix} \psi_+(\boldsymbol{x}) \\ \psi_-(\boldsymbol{x}) \end{pmatrix} \tag{B.46}$$

so that $|\psi_+(\boldsymbol{x})|^2$ is the probability density for finding the particle near $\boldsymbol{x}$ with a spin component of $+\frac{1}{2}\hbar$ along the quantization axis, and similarly for $\psi_-(\boldsymbol{x})$. The operator s^3 must be a diagonal 2×2 matrix with eigenvalues $\pm\frac{1}{2}\hbar$, and in fact the operators for the three spin components are $s^i = \frac{1}{2}\hbar\sigma^i$, where σ^i are the *Pauli matrices* shown in (7.28). Readers may readily verify that these matrices obey the commutation relations (B.31). The somewhat deeper understanding of spin that arises from the relativistic theory is discussed in chapter 7.

The existence of spin-$\frac{1}{2}$ particles requires us to enlarge our view of the rotation group. According to the general rule (B.9), the matrix $U(\boldsymbol{\alpha})$ which rearranges the components of a spin-$\frac{1}{2}$ wavefunction under a rotation is

$$U(\boldsymbol{\alpha}) = \exp(\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{s}/\hbar) = \exp(\tfrac{1}{2}\mathrm{i}\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}). \tag{B.47}$$

Now, the square of each σ^i is the unit 2×2 matrix and it is straightforward to show from a Taylor series similar to (B.7) that

$$U(\boldsymbol{\alpha}) = \cos(\tfrac{1}{2}\alpha) + \mathrm{i}\sin(\tfrac{1}{2}\alpha)\boldsymbol{n} \cdot \boldsymbol{\sigma}. \tag{B.48}$$

Evidently, for a rotation through an angle of 2π, we get $U = -1$, whereas the rotation of a vector through this angle using $R(\boldsymbol{\alpha})$ obviously leaves the vector unchanged. For spin-$\frac{1}{2}$ wavefunctions, any rotation angle between 0 and 4π leads to a distinct transformation. A rotation through an angle of $\alpha + 2\pi$ leaves the spin pointing in the same direction as a rotation through the angle α, but changes the sign of the wavefunction. Before discussing this further, it will be useful to know a little about the classification of Lie groups and their Lie algebras.

Mathematicians have achieved a complete classification of all the Lie groups. This is too complicated an enterprise for me to enter fully into it here, but one important ingredient is the question of what is left unchanged by the transformations that constitute the group. In the case of rotations of a 3-dimensional position vector, the length of a vector, $|\boldsymbol{r}|^2 = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}$ is unchanged. This means that

$$\boldsymbol{r}'^{\mathrm{T}}\boldsymbol{r}' = \boldsymbol{r}^{\mathrm{T}}R^{\mathrm{T}}(\boldsymbol{\alpha})R(\boldsymbol{\alpha})\boldsymbol{r} = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r} \tag{B.49}$$

for which we require $R(\boldsymbol{\alpha})$ to be an *orthogonal* matrix, $R^{\mathrm{T}}(\boldsymbol{\alpha}) = R^{-1}(\boldsymbol{\alpha})$. The group consisting of all real 3×3 orthogonal matrices is the *orthogonal group* O(3), though we have seen that it can be represented by matrices of other sizes as well. However, not all orthogonal matrices can be interpreted as rotations. For example, the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{B.50}$$

is an orthogonal matrix which reverses one component of a vector—an effect that cannot be achieved by a rotation. The rotations actually constitute the *special orthogonal group* SO(3) of 3×3 orthogonal matrices whose determinants are equal to 1. More generally, by considering real $N \times N$ orthogonal matrices, we arrive at the groups O(N) and SO(N).

Lorentz transformations of a 4-vector u^{μ} in Minkowski spacetime, $u^{\mu'} = \Lambda^{\mu'}{}_{\mu}u^{\mu}$, leave the scalar product $u_{\mu}u^{\mu}$ unchanged. In the notation we are using here, we can write this scalar product as $u^{\mathrm{T}}\eta u$, where η is the matrix (2.8), and we see that the matrix Λ must satisfy the condition

$$\Lambda^{\mathrm{T}}\eta\Lambda = \eta. \tag{B.51}$$

These matrices constitute the group O(3,1), or SO(3,1) if we restrict ourselves to matrices of unit determinant. More generally, we could substitute for η a diagonal $(p+q) \times (p+q)$ matrix with p of its diagonal elements equal to +1 and the remaining q elements equal to -1. To be definite, we take $p \geq q$; however the condition (B.51) is clearly the same if we replace η with $-\eta$, so we could also take p negative elements and q positive ones. Then the matrices Λ constitute the group O(p, q) or SO(p, q).

In quantum mechanics, a change of basis in the Hilbert space of state vectors is a *unitary* transformation (see exercise 5.6). To be definite, consider a wavefunction $\boldsymbol{\psi}$ such as (B.46) which is a complex column matrix with N components. In order to preserve the probabilistic interpretation, a transformed wavefunction $\boldsymbol{\psi}' = U\boldsymbol{\psi}$ must satisfy

$$\boldsymbol{\psi}'^{\dagger}\boldsymbol{\psi}' = \boldsymbol{\psi}^{\dagger}U^{\dagger}U\boldsymbol{\psi} = \boldsymbol{\psi}^{\dagger}\boldsymbol{\psi} \tag{B.52}$$

and U must be a unitary matrix, $U^{\dagger} = U^{-1}$. The set of unitary $N \times N$ matrices constitutes the group U(N), or SU(N) if we restrict them to have determinant equal to 1.

There are several other variations on the same theme. For example, if we replace η in (B.51) with the matrix (3.100) we arrive at the *symplectic* groups Sp(N).

It might seem that only the unitary groups U(N) and SU(N) are relevant to quantum-mechanical systems, but this is not so. What *is* necessary for a quantum-mechanical symmetry group is that at least some of its representations should be unitary, which means that the transformation matrices of these representations are unitary in addition to having the properties that specify the group. In the case of O(N), all the representations are unitary, because a real orthogonal matrix is a unitary matrix. However, the matrices (B.48), while they are unitary, are not real. In fact, the half-odd-integer representations of the rotation group do not belong to the group SO(3)—they belong to SU(2). It happens that the commutation relations of the generators of SO(3) and SU(2) are identical. The largest group with these commutations, called the *covering group*, is SU(2), which must be regarded as the full rotation group if we wish to include the half-odd-integer representations, as we must in any situation involving spin-$\frac{1}{2}$ particles.

Appendix C

Natural Units

When we deal with everyday physical situations, it is convenient to use the SI system of units, based upon the metre as a unit of length, the second as a unit of time and the kilogram as a unit of mass. For doing fundamental physics, it is usually much more convenient to use a system of units, known as *natural units*, in which the constants $\hbar$ and c are both equal to 1. Since three basic units need to be defined, this leaves us with one unit still to be chosen. In experiments which study the properties of fundamental particles, the quantity that is most easily controlled is the energy of a particle which has been accelerated by means of electromagnetic fields and whose charge is some multiple of the fundamental charge e, so a convenient choice for the remaining unit is some multiple of the electron-volt. To be definite, let us choose the MeV (10^6eV), which is approximately twice the rest energy of an electron. The conversion factors which allow us to change between SI and natural units are:

$$\begin{aligned}1\ \text{MeV} &= 1.602\,176\,462\times10^{-13}\,\text{J}\\ \hbar &= 1.054\,571\,596\times10^{-34}\,\text{J s} = 6.582\,118\,89\times10^{-22}\,\text{MeV s}\\ c &= 2.997\,924\,58\times10^{8}\,\text{m s}^{-1}\\ \hbar c &= 1.973\,269\,602\times10^{-13}\,\text{MeV m}.\end{aligned}$$

Thus, for example, if $t(\text{s})$ is a time interval measured in seconds, then $t(\text{MeV}^{-1}) = t(\text{s})/\hbar$ is the equivalent interval in natural units, where the unit of time is MeV^{-1}. Some useful conversions are:

$$\begin{aligned}&\text{time:} && t(\text{s}) = 6.582\,118\,89\times10^{-22}\,t(\text{MeV}^{-1})\\ &\text{distance:} && l(\text{m}) = 1.973\,269\,6\times10^{-13}\,l(\text{MeV}^{-1})\\ &\text{mass:} && m(\text{kg}) = 1.782\,661\,73\times10^{-30}\,m(\text{MeV}).\end{aligned}$$

From a theoretical point of view, the use of natural units is more than a matter of convenience: it embodies much of our understanding of the way the world is. If, for example, we measure the speed of sound in a particular material, it makes sense to ask why this speed has the particular value we measure. We can set about calculating it in terms of the density and elastic modulus of the material and these in turn depend on the masses of its constituent atoms and the forces that act on them. However, it makes no sense to ask the same question of the speed of light. According to the theories of relativity, the metrical structure of space and time implies that time intervals and distances are really things of the same kind, and there is no fundamental reason for measuring them in different units. The reason for the appearance of a fundamental 'velocity' c is just that we traditionally measure these two quantities relative to two different standards. The value $c = 1$ is, in every sense of the word, the *natural* value. The number $2.99\ldots \times 10^8$ does not represent the value of any genuine physical quantity. It is properly thought of as being merely a conversion factor that relates our procedures for calibrating rulers and clocks. (Whether electromagnetic radiation always travels through empty space with precisely this speed may be another matter, but all the indications are that it does.) There is, of course, a good reason for our using different standards for measurements of time and distance intervals. It is that our conscious experiences of these quantities are of quite different kinds. In the equations of theoretical physics, this obvious difference is represented by nothing more than the minus signs in (2.8). This leaves, in my view, deep unresolved questions about the relationship between the universe as described by physics and the actual perceptions of sentient beings such as physicists.

In a somewhat similar way, quantum theory tells us that the notions of energy and momentum are essentially equivalent to those of frequency and inverse wavelength. At an elementary level, this equivalence is manifest in the de Broglie relations (5.1) and (5.2). More fundamentally, it arises from the canonical commutation relations and the role of the energy and momentum operators as the generators of spacetime translations. The real significance of Planck's constant is not that, for example, the magnitude of the right-hand side of (5.38) is $1.054\ldots \times 10^{-34}$ J s, but simply that it is *not zero*. The fact that this commutator is non-zero means that there is a fundamental relationship between momenta and intervals of distance, and there is therefore no fundamental reason for measuring them in independent units. Thus, the *natural* way of measuring momentum is as an inverse length, and the constant $\hbar$ is a conversion factor which translates an inverse length into our traditional units of mass × velocity. Even though momentum is not something we perceive directly, it is fair to say that the notion of momentum as an inverse distance does not correspond in an obvious way to our ordinary experience of the behaviour of physical objects. As with time and distance, therefore, there is a good reason for our traditional momentum units. The fact that momentum does not ordinarily appear to us as an inverse wavelength is, in my view, one of the deep unresolved mysteries of the interpretation of quantum theory. Whether this mystery is also bound up with the place of sentient beings in the physical world, I am not sure.

From a theoretical point of view, the SI system of units treats electromagnetic quantities in a curious way. In my opinion, this creates deep mysteries where none actually exist! In a vacuum, the electrostatic potential energy of, say, two electrons treated as classical particles a distance r apart is

$$V(r) = e^2/4\pi\epsilon_0 r$$

where the quantity $\epsilon_0 = 8.854\,187\ldots \times 10^{-12}\,\mathrm{F\,m^{-1}}$ is called the *permittivity of free space*. The physical content of this is that the potential energy is proportional to $1/r$, with a constant of proportionality equal to $e^2/4\pi\epsilon_0$. This quantity, which measures the strength of electrical forces clearly has the dimensions of (energy × distance) and, in natural units, is equal to the *fine structure constant*

$$\alpha = e^2/4\pi\epsilon_0\hbar c = 7.297\,352\,533 \times 10^{-3} \approx 1/137$$

which is dimensionless. The factor of 4π in the denominator has a geometrical significance (see equation (9.84)), being the surface area of a unit sphere, but the constant ϵ_0 is merely a conversion factor which relates the SI unit of charge, the Coulomb, to the units of energy and distance. It cannot be emphasized too strongly that ϵ_0 *does not* refer to any physical property of the vacuum. Similarly, magnetic forces involve a quantity μ_0, called the *permeability of free space*, whose value is *defined* to be $4\pi \times 10^{-7}\,\mathrm{H\,m^{-1}}$. Since its value is defined, μ_0 also cannot refer to any physical property of the vacuum and it too is no more than a conversion factor. The product $\epsilon_0\mu_0$ is equal to $1/c^2$ which, as we have seen, is also a conversion factor in the relativistic view of the world. If, when dealing with electromagnetism in SI units, we were to measure all charges in units of $e/\sqrt{\epsilon_0}$, then only the constant c would ever appear. The reason why c appears is that the magnetic field generated by a moving charge is obtained by a Lorentz transformation of the electric field in its rest frame and, if the velocity of the charge is v, depends on v/c.

There is, therefore, no real need for an independent unit of electric charge. Classically, the strength of electromagnetic forces involving an SI charge q is measured in purely mechanical units by q^2/ϵ_0. Quantum-mechanically, the strength of electromagnetic forces between fundamental particles is measured by the dimensionless number α, although a proper characterization of this strength requires the running coupling constant discussed in chapter 9.

It might be wondered whether some third fundamental constant, in addition to $\hbar$ and c, should be used to define a system of natural units in which no arbitrary choice of a third unit would be called for. One possibility would be to take the mass of some fundamental particle as a basic unit. The trouble here is that there are many particles to choose from. At present, we do not properly understand the origin of particle masses and there is no good reason for regarding, say, the electron or muon as especially fundamental. It is quite possible to imagine a universe in which, although $\hbar$ and c had the same significance as in ours, there were no electrons or muons. The only serious candidate for a third truly fundamental constant is Newton's gravitational constant G. By using $\hbar$, c and G, we can construct three fundamental units of mass, length and time, which are the

Planck units

Planck time:	$(G\hbar c^{-5})^{1/2} = 5.389 \times 10^{-44}\,\mathrm{s}$
Planck length:	$(G\hbar c^{-3})^{1/2} = 1.615 \times 10^{-35}\,\mathrm{m}$
Planck mass:	$(\hbar c/G)^{1/2} = 2.176 \times 10^{-8}\,\mathrm{kg}.$

It is often also useful to define the *Planck energy* $(\hbar c^5/G)^{1/2} = 1.22 \times 10^{19}\,\mathrm{GeV}$. Unfortunately, it is not quite clear whether G has the same fundamental status as $\hbar$ and c. It appears, indeed, to be more like e, in that it measures the strength of gravitational forces. Whereas $\hbar$ and c are merely conversion factors, it seems possible to imagine that electromagnetism and gravity could have been either weaker or stronger than they actually are, and that in that sense e and G measure genuine physical properties of our particular world. In any system of units with $\hbar = c = 1$, e is properly measured by the dimensionless fine-structure constant (the strengths of the weak and strong interactions are measured by similar dimensionless constants) and cannot provide a third basic unit. G, on the other hand, cannot be combined with $\hbar$ and c to form a dimensionless measure of the strength of gravity. This fact, as discussed in chapter 12, is symptomatic of the difficulties we experience in trying to reconcile gravity with quantum mechanics, and might be an indication that G is not as fundamental as it appears. According to string theory (see chapter 15), the gravitational constant apparent to us is determined by a fundamental string tension α', through relations which also involve gauge couplings and parameters which characterize the compactification of a 10- or 11-dimensional spacetime.

Appendix D

Scattering Cross-Sections and Particle Decay Rates

When analyzing the results of a high-energy scattering experiment, we typically consider an initial state containing two particles, with 4-momenta k_1 and k_2, and wish to know the probability of obtaining a given final state containing, say, N particles with momenta $k'_1, \dots, k'_N$. Actually, the probability of a final state with *exactly* these momenta is generally zero, and we ask instead for the probability that the first final-state particle has its 3-momentum in the range $\mathrm{d}^3k'_1$ near $\boldsymbol{k}'_1$ and so on. These probabilities are conventionally expressed in terms of *cross-sections*, which can be understood picturesquely in the following way. We consider an incident particle, number 1, heading in the general direction of a stationary target particle, number 2. In the plane containing the target particle and perpendicular to the momentum of the incident particle, we draw an annulus surrounding the target particle of area $\mathrm{d}\sigma$, and imagine that any incident particle that passes through this annulus will give rise to the specified final state. The greater the probability of this event, the larger is the cross-section $\mathrm{d}\sigma$. This is not what actually happens—the picture simply gives a way of quantifying the probability in the following way. Suppose we have a beam of incident particles with a flux j equal to the number of particles crossing a unit cross-sectional area per unit time and a target containing n particles per unit volume. The number of scattering events per unit time per unit volume of the target is given by

$$\text{number of events/unit time/unit volume} = jn\,\mathrm{d}\sigma. \tag{D.1}$$

Regardless of our simple picture, this defines the differential cross-section $\mathrm{d}\sigma$.

The quantities we can attempt to calculate theoretically are S-matrix elements of the form $\langle k'_1, \dots, k'_N; \text{out}|k_1, k_2; \text{in}\rangle$. Since energy and momentum are conserved, this matrix element is proportional to $\delta(P_\mathrm{f} - P_\mathrm{i})$, where P_i and P_f are the total 4-momenta of the initial and final states. To be specific, we write it as

$$\langle k'_1, \dots, k'_N; \text{out}|k_1, k_2; \text{in}\rangle = (2\pi)^4\delta(P_\mathrm{f} - P_\mathrm{i})T_{\mathrm{fi}}. \tag{D.2}$$

According to (5.9), the probability we want is proportional to the square magnitude of this quantity, which involves the square of the δ function. In one of these two δ functions, we are entitled to set the argument to zero. This gives an infinite value, which can be interpreted in the following way. Using the representation (A.11), but remembering that the argument of our function is a 4-momentum, we have

$$(2\pi)^4\delta(0) = \int \mathrm{d}^3x\,\mathrm{d}t. \tag{D.3}$$

If we imagine observing a large target volume V for a long time T, we can interpret this spacetime integral as the product VT.

We must now adjust the basic probability formula (5.9) to take into account that our final-particle states are normalized according to (7.18) rather than (5.12). Consider, therefore, a single-particle state $|\Psi\rangle$ such that $\langle\Psi|\Psi\rangle = 1$. For a particle in this state, the probability of finding it to have a 3-momentum in the range d^3k near $\boldsymbol{k}$ must be of the form

$$P(k|\Psi)\mathrm{d}^3k = |\langle k|\Psi\rangle|^2 g(\boldsymbol{k})\mathrm{d}^3k = \langle\Psi|k\rangle\langle k|\Psi\rangle g(\boldsymbol{k})\mathrm{d}^3k \tag{D.4}$$

where the function $g(\boldsymbol{k})$ is chosen to ensure that $\int P(k|\Psi)\mathrm{d}^3k = 1$. This implies that

$$\int \mathrm{d}^3k\, g(\boldsymbol{k})|k\rangle\langle k| = \hat{I} \tag{D.5}$$

where $\hat{I}$ is the identity operator (see exercise 5.4). By acting with this operator on the vector $|k'\rangle$ and using (7.18), it is straightforward to see that $g(\boldsymbol{k}) = [(2\pi)^3 2\omega(\boldsymbol{k})]^{-1}$.

To calculate the average number of scattering events per unit time per unit volume which give rise to final states in the range that we specified at the outset, we thus take the squared magnitude of the matrix element (D.2), divide by VT (which equals $(2\pi)^4\delta(0)$) and multiply by the factor

$$\mathrm{d}\rho_{\mathrm{f}} = C_{\mathrm{f}} \prod_{i=1}^{N} \frac{\mathrm{d}^3k_i'}{(2\pi)^3 2\omega(\boldsymbol{k}_i')}. \tag{D.6}$$

This ‘phase space’ factor includes a factor of $g(\boldsymbol{k})\mathrm{d}^3k$ for each final-state particle. The number C_{f} is included to account for any sets of identical particles in the final state: for any set of n identical particles, C_{f} includes a factor of $1/n!$, because rearrangements of these particles do not count as distinct states. The quantity we arrive at in this way is the scattering rate per unit volume defined in (D.1)

$$jn\,\mathrm{d}\sigma = (2\pi)^4\delta(P_{\mathrm{f}} - P_{\mathrm{i}})|T_{\mathrm{fi}}|^2\mathrm{d}\rho_{\mathrm{f}} \tag{D.7}$$

provided that j and n are identified in accordance with the normalization of the initial particle states. As we saw in §7.2, this normalization implies that there are $2\omega(\boldsymbol{k})$ particles per unit volume. The target particles with mass m_2 are at rest, so

$n = 2m_2$, and if the incident particles are travelling with a speed v, then their flux is $j = 2\omega(\boldsymbol{k}_1)v$. Thus, our result for the cross-section is

$$\mathrm{d}\sigma = \frac{1}{4Q}(2\pi)^4\delta(P_\mathrm{f} - P_\mathrm{i})|T_\mathrm{fi}|^2\mathrm{d}\rho_\mathrm{f} \tag{D.8}$$

where $Q = \omega(\boldsymbol{k}_1)m_2 v$. This expression for Q is valid in the rest frame of the target particles, where the initial 4-momenta are $k_1 = (\omega(\boldsymbol{k}_1), \boldsymbol{k}_1)$ and $k_2 = (m_2, \mathbf{0})$. It is easily shown that $v = |\boldsymbol{k}_1|/\omega(\boldsymbol{k}_1)$ and that

$$Q = \left[(k_1 \cdot k_2)^2 - m_1^2 m_2^2\right]^{1/2}. \tag{D.9}$$

This is a Lorentz scalar, expressed in terms of the 4-vector momenta, and is therefore valid in any frame.

Although the quantities T_fi, $\mathrm{d}\rho_\mathrm{f}$ and Q all depend on the normalization of particle states, the differential cross-section $\mathrm{d}\sigma$ does not depend on this normalization, and can be compared directly with a cross-section derived from an experimental situation in which the density of target particles and the flux of incident particles may be quite different. In practice, it may be neither practical nor desirable to determine the energy, momentum and direction of every particle emerging from a high-energy collision. In the study of deep inelastic scattering (see §12.4), for example, one may ask for the probability that the energy of the emerging electron is between E' and $E' + \mathrm{d}E'$ and that its direction lies in an element of solid angle $\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$ containing the direction specified by the polar angles θ and ϕ. In this case, we can write the phase-space factor in the form $\mathrm{d}\rho_\mathrm{f} = \widetilde{\rho}_\mathrm{f}(E', \theta, \phi, \{K_i\})\mathrm{d}E'\,\mathrm{d}\Omega\prod_i \mathrm{d}K_i$, where the variables K_i account for all the unspecified momentum components of other final-state particles. The probability of finding the electron with its momentum in the specified range, regardless of the states of the other particles, is then measured by the differential cross-section

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega\,\mathrm{d}E'} = \int \frac{1}{4Q}(2\pi)^4\delta(P_\mathrm{f} - P_\mathrm{i})|T_\mathrm{fi}|^2\widetilde{\rho}_\mathrm{f}\prod_i \mathrm{d}K_i. \tag{D.10}$$

If some of the particles have spin, then the matrix element T_fi may depend on their spin polarization state. Depending on whether the initial particles are prepared with a definite polarization, and whether the polarizations of the final-state particles are determined by a given set of detectors, we may want to sum over the polarizations of final-state particles (so as to account for all the possibilities) and/or to average over those of the initial particles (so as to account for our ignorance of these details of the initial state).

In the same way, we can consider an initial state containing a single unstable particle, say of mass m, and work out the probability per unit time $\mathrm{d}\Gamma$ for it to decay into a final state specified as above. The result, valid in the rest frame of the decaying particle, is

$$\mathrm{d}\Gamma = \frac{1}{2m}\,(2\pi)^4\delta(P_\mathrm{f} - P_\mathrm{i})|T_\mathrm{fi}|^2\mathrm{d}\rho_\mathrm{f}. \tag{D.11}$$

By summing over all the decay modes (that is, over all the possible combinations of particles that might be produced) and integrating over the momenta of the emerging particles, we get the total decay probability per unit time Γ, and the lifetime of the particle is $1/\Gamma$.

Bibliography

I give here a list of textbooks and review articles on the various subjects covered during the Tour. Those marked (B) will be most useful for preliminary background reading. Those marked (E) offer accounts of a descriptive nature or at a modest technical level. A few, marked (A), are considerably more advanced than this one. The others, while treating their specialized subjects in more detail than I do, should be readily understood by anyone who has mastered the contents of this book.

General Theoretical Physics

Bederson B (ed) 1999 *Rev. Mod. Phys.* **71** (A collection of review articles surveying the state of Physics at the end of the second millennium.)

Longair M S 1984 *Theoretical Concepts in Physics* (Cambridge: Cambridge University Press) (B)

Classical Mechanics and Classical Electromagnetism

Goldstein H 1980 *Classical Mechanics* (Reading, MA: Addison Wesley) (B)

Jackson J D 1999 *Classical Electrodynamics* (New York: Wiley) (B)

Lorrain P, Corson D R and Lorrain F 1988 *Electromagnetic Fields and Waves* (New York: Freeman) (B)

Solymar L 1984 *Lectures on Electromagnetic Theory* (Oxford: Oxford University Press)

Geometry, Relativity, Gravitation and Cosmology

(See also under Mathematical Methods.)

Abbott L F and Pi S-Y (eds) 1986 *Inflationary Cosmology* (Singapore: World Scientific) (A)

Brandenberger R H 1985 Quantum field theory methods and inflationary universe models *Rev. Mod. Phys.* **57** 1

Coles P and Lucchin F 1995 *Cosmology: The Origin and Evolution of Cosmic Structure* (Chichester: Wiley)

Foster J and Nightingale J D 1995 *A Short Course in General Relativity* (New York: Springer-Verlag)

Hawking S W and Ellis G F R 1973 *The Large-Scale Structure of Spacetime* (Cambridge: Cambridge University Press) (A)

Kenyon I 1990 *General Relativity* (Oxford: Oxford University Press)

Liddle A R and Lyth D H 2000 *Cosmological Inflation and Large-Scale Structure* (Cambridge: Cambridge University Press)

Kolb E W and Turner M S 1990 *The Early Universe* (Redwood City, CA: Addison-Wesley)

Lightman A P, Press W H, Price R H and Teukolsky S A 1979 *Problem Book in Relativity and Gravitation* (Princeton, NJ: Princeton University Press)

Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco: Freeman)

Peebles P J E 1971 *Physical Cosmology* (Princeton, NJ: Princeton University Press)

Peebles P J E 1993 *Principles of Physical Cosmology* (Princeton, NJ: Princeton University Press)

Schutz B F 1985 *A First Course in General Relativity* (Cambridge: Cambridge University Press)

Silk J 2001 *The Big Bang* (San Francisco: Freeman) (E)

Wald R M 1984 *General Relativity* (Chicago: Chicago University Press)

Weinberg S 1972 *Gravitation and Cosmology* (New York: Wiley)

Weinberg S 1993 *The First Three Minutes* (London: Fontana) (E)

Quantum Theory, Quantum Field Theory and Elementary Particles

Aitchison I J R and Hey A J G 1989 *Gauge Theories in Particle Physics* (Bristol: Institute of Physics Publishing)

Barnett R M *et al* 1996 Particle Physics Summary *Rev. Mod. Phys.* **68** 611

Birrell N D and Davies P C W 1982 *Quantum Fields in Curved Space* (Cambridge: Cambridge University Press)

Cheng T-P and Li L-F 1984 *Gauge Theories of Elementary Particle Physics* (Oxford: Oxford University Press)

Coleman S 1985 *Aspects of Symmetry* (Cambridge: Cambridge University Press) (A)

Collins J C 1985 *Renormalization* (Cambridge: Cambridge University Press) (A)

Donoghue J F, Golowich E and Holstein B R 1994 *Dynamics of the Standard Model* (Cambridge: Cambridge University Press)

Groom D E *et al* 2000 Review of Particle Physics *The European Physical Journal* **C15** 1 (`http://pdg.lbl.gov`)

Halzen F and Martin A D 1984 *Quarks and Leptons: An Introductory Course in Modern Particle Physics* (New York: Wiley)

Itzykson C and Zuber J-B 1985 *Quantum Field Theory* (New York: McGraw-Hill) (A)

Perkins D H 2000 *Introduction to High-Energy Physics* (Cambridge: Cambridge University Press)

Rajaraman R 1987 *Solitons and Instantons* (Amsterdam: North-Holland)

Ramond P 1990 *Field Theory: A Modern Primer* (Redwood City, CA: Addison-Wesley)

Ryder L H 1996 *Quantum Field Theory* (Cambridge: Cambridge University Press)

Schiff L I 1968 *Quantum Mechanics* (New York: McGraw-Hill)

Sudbery A 1986 *Quantum Mechanics and the Particles of Nature* (Cambridge: Cambridge University Press)

Taylor J C 1976 *Gauge Theories of Weak Interactions* (Cambridge: Cambridge University Press)

Ticciati R 1999 *Quantum Field Theory for Mathematicians* (Cambridge: Cambridge University Press)

Vilenkin A and Shellard E P S 2000 *Cosmic Strings and Other Topological Defects* (Cambridge: Cambridge University Press)

Weinberg S 1996 *The Quantum Theory of Fields* Vols. 1 and 2 (Cambridge: Cambridge University Press) (A)

Weinberg S 2000 *The Quantum Theory of Fields* Vol. 3 (Cambridge: Cambridge University Press) (A)

Zinn-Justin J 1996 *Quantum Field Theory and Critical Phenomena* (Oxford: Oxford University Press) (A)

Thermodynamics, Statistical Mechanics and Phase Transitions

Amit D J 1984 *Field Theory, the Renormalization Group and Critical Phenomena* (Singapore: World Scientific)

Dalvit D A R, Frastai J and Lawrie I D 1999 *Problems on Statistical Mechanics* (Bristol: Institute of Physics Publishing)

Fetter A L and Walecka J D 1971 *Quantum Theory of Many-Particle Systems* (New York: McGraw-Hill)

Goldenfeld N 1992 *Lectures on Phase Transitions and the Renormalization Group* (Reading, MA: Addison-Wesley)

Huang K 1987 *Statistical Mechanics* (New York: Wiley)

Pippard A B 1966 *The Elements of Classical Thermodynamics* (Cambridge: Cambridge University Press) (B)

Reichl L E 1998 *A Modern Course in Statistical Physics* (New York: Wiley)

Tinkham M 1996 *Introduction to Superconductivity* (New York: McGraw-Hill)

String Theory

Duff M J (ed) 1999 *The World in Eleven Dimensions* (Bristol: Institute of Physics Publishing) (A)

Green M B, Schwarz J H and Witten E 1988 *Superstring Theory* (Cambridge: Cambridge University Press) (A)

Greene B 2000 *The Elegant Universe* (London: Vintage) (E)

Polchinsky J 1998 *String Theory* (Cambridge: Cambridge University Press) (A)

Mathematical Methods

Cornwell J F 1984 *Group Theory in Physics* (London: Academic Press)

de Azcárraga J A and Izquierdo J M 1995 *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics* (Cambridge: Cambridge University Press) (A)

Jones H F 1998 *Groups, Representations and Physics* (Bristol: Institute of Physics Publishing)

Nakahara M 1990 *Geometry, Topology and Physics* (Bristol: Institute of Physics Publishing) (A)

Schutz B F 1980 *Geometrical Methods of Mathematical Physics* (Cambridge: Cambridge University Press)

Simmons G F 1963 *Introduction to Topology and Modern Analysis* (Tokyo: McGraw-Hill)

Tung W-K 1985 *Group Theory in Physics* (Singapore: World Scientific)

References

Abe F *et al* 1989 *Phys. Rev. Lett.* **63** 720

Abrams G F *et al* 1989 *Phys. Rev. Lett.* **63** 724

Adeva B *et al* 1989 *Phys. Lett.* B **231** 509

Albrecht A, Ferriera P, Joyce M and Prokopec T 1994 *Phys. Rev.* D **50** 4807

Albrecht A and Steinhardt P J 1982 *Phys. Rev. Lett.* **48** 1220

Anderson C D 1933 *Phys. Rev.* **43** 491

Barrow J D 1983 *Fundam. Cosmic Phys.* **8** 83

Bernstein J, Brown L S and Feinberg G 1989 *Rev. Mod. Phys.* **61** 25

Block N, Flanagan O and Güzeldere G (eds) 1997 *The Nature of Consciousness* (Cambridge, MA: MIT Press)

Boyanovsky D, Cormier D, de Vega H J, Holman R and Kumar S P 1998 *Phys. Rev.* D **57** 2166

Burles S and Tytler D 1998 *Astrophys. J.* **499** 699; **507** 732

Cardy J L 1987 in *Phase Transitions and Critical Phenomena* vol 11, ed C Domb and J L Lebowitz (London: Academic)

Decamp D *et al* 1989 *Phys. Lett.* B **231** 519

Dirac P A M 1928 *Proc. R. Soc.* A **117** 610

——1929 *Proc. R. Soc.* A **126** 360

Domb C and Green M S (eds) 1976 *Phase Transitions and Critical Phenomena* vol 6 (London: Academic)

Eddington A S 1929 *Space, Time and Gravitation* (Cambridge: Cambridge University Press)

Efetov K 1997 *Supersymmetry in Disorder and Chaos* (Cambridge: Cambridge University Press)

Einstein A 1905 *Ann. Phys., Lpz.* **17** 891, **18** 639

Evans M and McCarthy G G 1985 *Phys. Rev.* D **31** 1799

Georgi H and Glashow S L 1974 *Phys. Rev. Lett.* **32** 438

Giveon A and Kutasov D 1999 *Rev. Mod. Phys.* **71** 983

Glashow S L 1961 *Nucl. Phys.* **22** 579

Guth A 1981 *Phys. Rev.* D **23** 347

Guth A and Pi S-Y 1982 *Phys. Rev. Lett.* **49** 1110

——1985 *Phys. Rev.* D **32** 1899

Hawking S W 1974 *Nature* **248** 30

Intriligator K and Seiberg N 1996 *Nucl. Phys. B Proc. Suppl.* **45** 1

Jackiw R 1977 *Rev. Mod. Phys.* **49** 681

Kaluza T 1921 *Sitzungsber. Preuss. Acad. Wiss.* 966

Kibble T W B 1976 *J. Phys. A: Math. Gen* **9** 1387

Kirzhnits D A and Linde A D 1972 *Phys. Lett.* B **42** 471

Klein O 1926 *Z. Phys.* **37** 985

Landsberg P T (ed) 1982 *The Enigma of Time* (Bristol: Adam Hilger)

Lawrie I D 1988 *Nucl. Phys.* B **301** 685

——1999 *Phys. Rev.* D **60** 063510

Lawrie I D and Epp R J 1996 *Phys. Rev.* D **53** 7336

Lawrie I D and Lowe M J 1981 *J. Phys. A: Math. Gen.* **14** 981

Lawrie I D and Sarbach S 1984 *Phase Transitions and Critical Phenomena* vol 9, ed C Domb and J L Lebowitz (London: Academic)

Linde A D 1982 *Phys. Lett.* B **108** 389

——1983 *Phys. Lett.* B **129** 177

Lockwood M 1989 *Mind, Brain and the Quantum* (Oxford: Blackwell)

Lorentz H A 1904 *Proc. Acad. Sci. Amsterdam* **6** 809

Lucas J R 1973 *A Treatise on Space and Time* (London: Methuen)

Mandelstam S 1975 *Phys. Rev.* D **11** 3026

Maxwell J C 1864 *Phil. Trans. R. Soc.* **155** 459

——1873 *A Treatise on Electricity and Magnetism* (Oxford: Clarendon) reprinted in 1954 (New York: Dover)

Mazenko G, Unruh W G and Wald R M 1985 *Phys. Rev.* D **31** 273

Mermin N D and Wagner H 1966 *Phys. Rev. Lett.* **17** 1133

Michelson A A and Morley E W 1887 *Am. J. Sci.* **34** 333 and *Phil. Mag.* **24** 449

Minkowski H 1908 Address to the 80th Assembly of German Natural Scientists and Physicians; translation in *The Principle of Relativity* (Methuen 1923) reprinted in 1952 (New York: Dover)

Morris R 1986 *Time's Arrows* (New York: Simon and Schuster)

Newton I 1686 *Philosophiae Naturalis Principia Mathematica* English translation by A Motte 1927. Revised translation ed F Cajori 1966 (Berkeley, CA and Los Angeles, CA: University of California Press)

Nienhuis B 1987 *Phase Transitions and Critical Phenomena* vol 11, ed C Domb and J L Lebowitz (London: Academic)

Onsager L 1944 *Phys. Rev.* **65** 117

Ornstein R E 1969 *On the Experience of Time* (Harmondsworth: Penguin)

Penzias A A and Wilson R W 1965 *Astrophys. J.* **142** 419

Perlmutter S *et al* 1998 *Nature* **391** 51

Pound R V and Rebka G A 1960 *Phys. Rev. Lett.* **4** 337

Prigogine I 1980 *From Being to Becoming* (San Francisco, CA: Freeman)

Salam A 1968 *Elementary Particle Physics (Nobel Symposium No 8)* ed N Svartholm (Stockholm: Almqvist and Wilsell)

Salam A and Ward J C 1964 *Phys. Lett.* **13** 168

Samuel S 1978 *Phys. Rev.* D **18** 1916

Schramm D N and Turner M S 1998 *Rev. Mod. Phys.* **70** 303

Schwarzschild K 1916 *Sitzungsber. Preuss. Acad. Wiss.* 189

Shapiro I I 1964 *Phys. Rev. Lett.* **13** 789

Smart J J C (ed) 1964 *Problems of Space and Time* (New York: Macmillan)

Starobinski A A 1982 *Phys. Lett.* B **117** 175

't Hooft G 1971 *Nucl. Phys.* B **33** 173, B **35** 167

——1974 *Nucl. Phys.* B **79** 276

Weinberg S 1967 *Phys. Rev. Lett.* **19** 1264
Wess J and Zumino B 1974 *Nucl. Phys.* B **70** 39
Whitrow G J 1975 *The Nature of Time* (Harmondsworth: Penguin)
Wilson K G and Fisher M E 1972 *Phys. Rev. Lett.* **28** 240
Yang C N 1952 *Phys. Rev.* **85** 809
Yang C N and Mills R L 1954 *Phys. Rev.* **96** 191

Index